表構造における意味的関係に基づくWWW検索性能の向上

松本　章代[†,††a)]　小西　達裕[†]　高木　朗[†††]　小山　照夫[††††]
三宅　芳雄[†††††]　伊東　幸宏[†]

Improvement in Performance of WWW Search Engines Based on Semantic Relation in Table Structure

Akiyo MATSUMOTO[†,††a)], Tatsuhiro KONISHI[†], Akira TAKAGI[†††], Teruo KOYAMA[††††], Yoshio MIYAKE[†††††], and Yukihiro ITOH[†]

あらまし　ウェブ検索エンジンに，ユーザが検索キーワードとして二つの語を入力した場合に，その2語が意味的関係をもって文書中に出現しているか否かを判定することにより，ウェブ検索エンジンの性能を向上させる手法を提案する．キーワード間の意味的関係を表現する構造として，本研究では表構造を取り上げる．提案手法を評価するために，既存の検索エンジンのフィルタリングツールを構築し，自作の評価用データセットを用いて実験を行った．実験の結果，今回提案した表構造を利用したランキング手法を追加することによって，これまでの文構造・見出し構造によるランキング手法と比較し，精度をほぼ下げることなく再現率を約12%向上させることを示した．検索者の求める情報が表構造で整理されていることが予想できるような検索キーワードの場合には，この表構造・文構造・見出し構造を合わせたランキング手法によって，もとにした既存の検索エンジンより上位の適合ページ数が増え，フィルタリングツールとして有効であることが認められた．

キーワード　表構造解析，情報検索，WWW，係り受け関係，自然言語処理

1. まえがき

ウェブ検索エンジンは日常的に広く用いられているが，現状では，まだ不適合ページを相当程度含む結果となることが少なくなく，決して満足のいくレベルとはいいがたい．したがってこの不適合ページを減らすことが一つの問題である．

検索エンジンのランキングについての従来の基本的なアプローチは，語の出現に関する（TF-IDFなどの）統計量や語間の距離などに基づくものが主流であった．そのような方法で誤検出されるページには，確かに指定されたキーワードは存在するものの，各々のキーワードが全く異なった文脈の中で独立して用いられており，結果的に検索意図を満たさないページであっても拾い上げられてしまうというケースが多く見受けられる．

そこで我々は，複数の検索キーワードが文書内においてどのような構造で結び付いているかに着目して，検索性能を向上させることを試みてきた．検索キーワードとして選択される語は，単に出現確率に基づいて選択されるのではなく，何らかの意味的な関係にある語が選択される傾向にあると考えられる．したがって，それらの語が文書中に同時に存在するということだけでなく，それらの語が意味的関係を表現し得る構造中に含まれる文書を特定することにより，検索性能の向上が期待できる．そこで，キーワード間の意味的関係を表す構造として文構造に着目し，キーワードが文書内で文構造によって結び付けられている文書を選択するフィルタを提案した[1]．その結果，もとにした検索エンジンの上位100件までを全体集合とした中で精度の向上を実現し，適合ページをより上位に集中さ

† 静岡大学，浜松市
Shizuoka University, Hamamatsu-shi, 432–8561 Japan
†† 東京工業高等専門学校，八王子市
Tokyo National College of Technology, Hachioji-shi, 193–0997 Japan
††† 言語情報処理研究所，小金井市
NLP Research Laboratory, Koganei-shi, 184–0014 Japan
†††† 国立情報学研究所，東京都
National Institute of Informatics, Tokyo, 101–8430 Japan
††††† 中京大学，豊田市
Chukyo University, Toyota-shi, 470–0393 Japan
a) E-mail: riir@inf.shizuoka.ac.jp

せることができた．

しかしながら一方で，検索キーワードによってはあまり精度が向上しない場合があること，再現率の面で改善が必要であることが判明した．これは，文献[1]では，キーワードが文書内において文構造で結び付いている場合のみを対象としていることに起因していると考えられる．

文書内でキーワードの意味的関係を表現する構造として，文構造以外にも表構造や見出し構造などがある．そこで我々は，文構造に加え，それらの構造でキーワード間の意味的関係が表現されている文書も拾い上げる方法を検討してきた．このうち見出し構造を利用した検索性能の向上については，文献[2]で報告した．本論文では表構造の利用について述べる．具体的には，表の構成要素間に意味的関係を認めることができる組合せを考え，意味的関係が認められる構成要素にキーワードが含まれている場合に，そのページを適合ページと判定することにより，検索性能の向上を図る．

なお，本論文では，検索クエリとして，2語のキーワードが与えられる状況を想定し，議論を行う．これは，

- キーワード間の係り受け関係は2語間で規定されるため，3語以上のキーワードを考える場合でも，それらが構成する係り受け構造は2語の係り受け構造の組合せとしてとらえることができること，
- Jansenら[3]や風間ら[4]の報告のように実際のウェブ検索エンジンにおいて1語ないし2語で検索されるケースが圧倒的に多いこと，

による．

係り受け構造で関係付けられている2語の意味的関係の一つとして，並列（同格）の関係を考えることができるが，本論文ではこれを取り扱わない．本論文では，一方の語が連体助詞や述語を伴って連体修飾句（節）を構成して他方を修飾する関係に限定して検討する．なお，我々が行った予備調査の結果においては，検索で用いられた2語のキーワードが，修飾–被修飾の関係にあるケース86%，並列（同格）の関係にあるケース5%，その他9%(注1)であった．

本論文では，**2.**で表のシンタックスとセマンティクスに関する考察を述べ，**3.**で表内に含まれるキーワードの位置関係から意味的関係の有無を判定する判定基準を検討する．**4.**では，文書中から表を抽出する方法について述べ，**5.**でシステムの構成を述べる．また，**6.**では評価実験の結果を紹介し，今回提案する表構造で表現された意味的関係をチェックする手法を追加することによって，これまでの文構造・見出し構造による判定と比較し，文献[1]の精度をほぼ維持しつつ再現率を大きく向上させることができること，また，検索結果として表をイメージさせるキーワードの場合に限定すると，上位20件中の適合ページ数，精度及びMAPの各々において高い性能が達成できることを示す．

2. 表の基本構造

2.1 HTML文書中に存在する表の利用

HTML文書には，数多くの表が含まれている．一般的に，表は情報を凝縮し，分かりやすく整理した形式にしたものである．このような表構造は，情報抽出や情報検索の分野において無視できないリソースである．

実際，表構造中から情報を抽出する手法については様々な手法が提案されている[5]．またウェブページ内の表に含まれる情報を抽出する研究も多数行われており，そのために表構造を解析する研究には大谷ら[6]，大前ら[7]，吉田ら[8]，板井ら[9]によるものなどがある．これらは，属性，属性値が書かれている位置を特定し属性と属性値の組合せを抽出する手法を提案するものである．例えば大谷らは，抽出した属性と属性値の情報によって知識DBを作成し，それを問合せシステム等に応用することを提案している．また，ウェブページ中の表の内容からその主題を推定する手法には佐藤ら[10]によるものなどがある．このように，ウェブページに含まれる表を解析しその情報を利用しようという研究は多数存在する．

一方，ウェブ検索エンジンにおいては，現在まで表構造を活用している状況には至っていない．従来のウェブ検索エンジンは，表内の関係を示すタグを取り除き，各セルの内容を単にテキストとして取り扱っている．そのため，表内に明示されている各セル間の関係を検索に反映させることができていない[11]．この問題を解消するために岩口らは，各セル間の関係を検索に反映させることを目的として，ウェブ空間上に存在する複雑な表構造を対象にし，表構造内の関係を保持したまま各セルの内容を索引化する手法を提案して

(注1)：その他のケースに分類されたのは，単語の出現頻度を考慮してキーワードを選択したと思われるもの（第1のキーワードが多義語であり，検索者が意図しない方のページを排除するため，『同じ分野の文書に出現していると思われる特徴的な語』を第2のキーワードとして付け加えたと思われるものなど）である．

いる[11]．これにより，セル中に存在している情報が，どのような見出しの行または列に存在しているものか等，表中での位置関係情報が利用可能となった．しかし，この研究では索引化手法の提案にとどまっており，直接検索性能の向上を図るところまでは議論されていない．これは，表の中における検索キーワードの配置をどのように評価すればよいかが明らかではなかったためであると考えられる．

2.2 検索精度の向上を目的とした表の利用

そこで本論文では，表の中における検索キーワードの配置とその評価の関係を整理して検索精度の向上を図る．そのためにまず，表の中の検索キーワードがどのような位置関係にあるとき意味的関係をどう評価すれば良いかを明らかにするため，表のシンタックスとセマンティクスに基づいて仮説を立てる．それを実データによって統計的に検証し，検索精度の向上に有効なものを抽出する．そして最終的に，表の中のキーワードの出現位置に応じて適合文書かどうかの判断を行う機能を備えた検索システムの実現を目指す．

そのためにまず，表のシンタックスとセマンティクスについて検討する．表とは共通な構造をもつ情報を集めて圧縮したものと考えることができる．一方，その構造は，見出しを除くと二次元のマトリックスの構造でしかない．このような構造で複数の均一な情報を表現する場合，個々の情報の表現と集積方法については，基本的には可能なパターンは以下の三つとなる．

(1) 1行で一つの情報を表現し，それらを複数行重ねて表を構成するパターン
(2) 1列で一つの情報を表現し，それらを複数列重ねて表を構成するパターン
(3) 1セルで一つの情報を表現し，それらを縦横2方向に集積して表を構成するパターン

このうち(1)と(2)は，行と列を入れ換えるだけで相互に変換可能であるという観点から同質の表の構成方法とみなせる．そこで(1)(2)の行（列）を基本単位とするタイプ（以下タイプ(i)）と(3)のセルを基本単位とするタイプ（以下タイプ(ii)）の2通りについて，詳細に検討する．

2.3 行（列）を基本単位とする表

まずタイプ(i)の表について，行が基本単位となっている例を図1に示し，以下項目ごとに述べる．列が基本単位となっている表も，行と列を読み替えるだけで同等である．

主なスクリプト言語

	開発者	発表年	実行速度
Perl	Larry Wall	1987	○
Python	Guido van Rossum	1995	○
Ruby	まつもとゆきひろ	1995	△

図1 タイプ(i)
Fig. 1 Type(i).

- 基本単位

一つの情報（実体，現象あるいはそれらの組合せ）を1行で構成する．

- 解釈方法

各項目の間の関係を表す一定の表現を補間することにより解釈できる．補間する表現はすべての行に対し共通である．例えば，言語名・開発者・発表年・実行速度が並んだ表（図1）の場合，すべての行について「○○は△△氏が□□年に開発したスクリプト言語であり，その実行速度は◇◇である.」と言い表すことが可能である．

- 構成

行ごとに同種の情報が格納されるので，列見出し以外は縦に同じクラスの情報が並ぶ．横方向はそれぞれ異なる．

- 見出し

列見出しは各列に格納される情報のクラスや，行全体で表現される情報の中での役割などを表すものが選ばれる．1行に相当する一つ（場合によっては複数）の実体/現象の属性が見出しとなり，属性値が内容セルにかかわることが多い．行見出しは，ない場合もある．行全体で表す情報を1語で表すことのできる名前が存在する場合，それが行見出しとなることが多い．属性やキーをグループ化できる場合は，見出しの階層構造化が行われる．

- 表全体のタイトル

実体/現象がレコードの数ぶん束ねられたものが表となっており，この場合の表のタイトルは「束ねられている実体/現象が何なのか（どういう実体/現象を束ねたものか）」を表すとみなせる．図1でいえば，束ねられている「Perl」「Python」「Ruby」はいずれも「主なスクリプト言語」である．

2.4 セルを基本単位とする表

続いてタイプ(ii)の表について，例を図2に示し，タイプ(i)同様，項目ごとに述べる．

- 基本単位

一つの情報を1セルで構成する．ただしこの場合，

郵便物の料金

	25gまで	50gまで	100gまで
定形郵便物	80円	90円	−
定形外郵便物	−	120円	140円

図 2　タイプ (ii)
Fig. 2　Type(ii).

1 セルだけで個別の情報を完全に表現できる場合は，二次元構造に集積させる必要はなく，箇条書きと同等である．あえて二次元構造にするのは 1 セルの解釈に二次元に構造化することによって付加可能な情報を利用可能とするためであると考えられる．そのような情報として，行及び列ごとに付加される「見出し」がある．「見出し」がそのような働きをもつとすると，各セルについては，同一行及び同一列の見出しと合わせて個別の情報の解釈を行う必要がある．ここでは，タイプ (ii) の表として，見出しを解釈に利用するタイプのみを対象とする．

- 解釈方法

行見出し・列見出しとセルの情報の間を補間して解釈できる．補完する表現はすべてのセルに対し共通である．例えば，郵便物の料金表（図 2）において，各行の見出しが郵便物の種別，各列の見出しが重量，セルに料金が記述されている場合，(見出し以外の) すべてのセルについて「郵便物の種別が△△郵便物であり，その重さが○ g までである場合，その料金は□□円である」と言い表すことができる．

- 構成

各セルに同種の情報が格納される．見出し以外は縦横両方向に同じクラスの情報が並ぶ．

- 見出し

「基本単位」で述べたように，行見出し・列見出しは，それらの行・列が交わるところの内容セルとともに個別の情報を構成する要素となる．見出しと内容セルの間にクラス-インスタンスのような関係はない．行見出し同士と列見出し同士は同じクラスの語となる．見出しは，内容セルに書かれた情報が真となる条件とみなすことができる．条件が三つ以上の場合は，見出しの階層構造化が起こる．例えば「速達かどうか」という条件が加わった場合は，図 3 のようになる．

- 表全体のタイトル

「条件 1 かつ条件 2 であるとき○○である」といった現象が（見出し以外の）セルの数分束ねられたものが表となっており，この場合の表のタイトルは「条件の組合せで何を規定しているのか」を示すとみなせる．図 2 では，郵便物の種別と重さで料金を規定している．

	25gまで		50gまで		100gまで	
	普通	速達	普通	速達	普通	速達
郵便物	80円	350円	90円	360円	−	−

図 3　見出しの階層構造化
Fig. 3　Hierarchical structure of headline.

3. 表構造における係り受けパターン

表構造は，**2.2** で説明したように，本来複数の文章で記述される内容を形式化し，一様な構造で表したものであり，表内の見出し，行見出し，列見出し，セル等の構成要素間に，一定の係り受け構造を含んでいる．そのような係り受け関係をもつ表の構成要素の中の二つの要素にキーワードが一致したということは，いわば，表の内容を記述する文の係り受け構造中の二つの語と，キーワードが一致したことに等しい．したがって，キーワード間に想定される（検索意図に即した）係り受け関係と，表の構成要素間に想定される係り受け関係とが一致する場合，係り受け関係にあるキーワードから想定される検索対象に関連する情報がその表内に書かれていると考えることができ，その表を含むページを適合ページとみなすことができる．すなわち，表を構成する係り受け関係の性質を，適合/不適合の判定に利用することができる．そこで本章では，表の構成要素間に意味的関係を認めることができる組合せから推定される「係り受けパターン」を検討する．ただしここでは，表の構成要素間の関係を，それに含まれるキーワードの関係と近似して議論を進める（この近似の限界については，**3.1** の末尾で述べる）．

3.1 では，表の構成要素間で意味的関係を認めることが可能なパターンを整理する．次いで **3.2** で，上述の近似を用いた場合にそれらのパターンの中で有効なものとそうでないものとを事例に基づいて検討する．

3.1 表の構成要素間の関係の検討

表の構成要素を

- 表見出し　表全体を支配する見出し
- 列見出し　各列の見出し
- 行見出し　各行の見出し
- 内容セル　表/列/行いずれの見出しでもないセル

の四つととらえたとき，上述のように，各構成要素間の意味関係を一様な形の文で表現可能である．ということは，それぞれの構成要素間に一定の意味的関係

(修飾–被修飾関係) が存在すると考えることができることを示している.

これらの構成要素のいずれかに二つの検索キーワードがそれぞれ出現する場合を想定すると, 可能な組合せは以下のとおりである.

(A) 表見出しと行 (または列) 見出し
(B) 表見出しと内容セル
(C) 行見出しと列見出し
(D) 行 (または列) 見出しと内容セル
(E) 同一見出し内または同一内容セル内
(F) 行 (または列) 見出し同士 (別々のセル)
(G) 異なる内容セル

この範囲内において, 二つの検索キーワード間に想定し得る検索意図と関連がある可能性がある場合は, その組合せを検討対象とする.

また, 表と意味的関係をもち得る表外の要素として,

- ページタイトル
- 段落の見出し
- 表を引用している文

などがある. 本論文では, これらについては, 処理可能で, かつ, 有効なものに限定して処理対象とすることにする. 現段階ではページタイトルと表の構成要素間の関係のみを取り扱うこととし, ページタイトルを表見出しと同様の扱いとすることにしている(注2). すなわち, 表見出しをページタイトルに置き換えたものを (A*)(B*) とし, それぞれ (A)(B) と同等に検討する. また, ページタイトルと表見出し間についても考慮すべきと考え,

(H) ページタイトルと表見出し

を加える.

①タイプ (i) の表において (A)〜(H) に挙げた各構成要素の組合せ間に意味的な関係が成立するか否か, するとしたらどのような関係なのかを整理し, ②タイプ (ii) の場合はどうか, ③表構造における係り受けパターンとして抽出すべきかどうか, について検討を行う. なお, ここでは説明の便宜上, タイプ (i) の表は「行」を基本単位としているものとして述べる.「列」を基本単位とする表の場合も, 行見出しと列見出しを入れ換えればよい.

(A) 表見出しと行 (または列) 見出し

① タイプ (i) における表見出しと列見出しの関係は, 束ねられている実体/現象とその属性であり, その属性値が検索意図を満たす. 例えば検索キーワードが「スクリプト言語　開発者」で,「スクリプト言語」が表見出し,「開発者」が列見出しに存在しているとき, 検索意図を満たす具体的な開発者の名前が, 該当列に列挙されているはずである. 一方, (タイプ (i) に行見出しがある場合) 表見出しと行見出しの関係は, 束ねられている実体/現象とそのうちの一つを特定するキーと考えられる. 例えば, 図 2 における「スクリプト言語」と「Python」の組合せであり,「スクリプト言語」の一つである「Python」に関連する情報が同じ行に存在していると思われる.

② タイプ (ii) において表見出しは, 各セルの意味を総称したものであるのに対し, 行/列見出しはセルごとの個別性をもたらす条件を表す. したがって, 表見出しと行/列見出しにそれぞれ検索キーワードが存在する場合は, 行/列見出しで指定される条件下での見出しに総称される事物や値を求めていると考えることができる. その場合, 行/列見出しの存在する行/列に求める情報があると推定される. 例えば, 表見出し「郵便物の料金」を構成する語とその条件として郵便物の種類や重量が具体的に検索キーワードとして指定された場合 (「定形外郵便　料金」など), その条件に該当する情報こそ検索者が求める情報と考えられる.

③ タイプ (i) の表見出しは行/列見出しと修飾–被修飾関係 (「スクリプト言語の Perl」「スクリプト言語の開発者」など) にあり, タイプ (ii) の表見出しは行/列見出しによって限定される関係 (「定形郵便物」の「郵便物の料金」は「80 円」か「90 円」) である. どちらも強い関係性があり, 表の中に求める情報が記述されていると考えられる. 検索キーワードの出現パターンとしても想定の範囲内であることから, 表構造における係り受けパターンとして抽出する. (A*) についても同様である.

(B) 表見出しと内容セル

① タイプ (i) においては, 表見出しと内容セルは, 見出しが表す集合概念と, それに含まれるある実体/現象の属性値の関係にある. この組合せによる検索は, その属性値をもつ実体/現象についての詳細な情報を求める場合に起こり得る. 例えば「スクリプト言語」が表見出し, Ruby の属性値である「まつもとゆきひろ」が内容セルに存在しているとき, 検索意図

(注2):(章・節・段落などで構造化されているページにおいて)「段落の見出し」と, その段落に含まれる表との修飾–被修飾関係については, 見出し構造の解析が必要であり, 階層化された見出し構造を用いた関係表現の一部として取り扱う予定である. 後述の評価における「見出し判定」には段落の見出しと表との修飾–被修飾関係を扱う処理は含まれていない.

を満たす「(スクリプト言語の開発者である) まつもとゆきひろ氏がどのような言語を開発したかに関する情報」は該当行に，また「まつもとゆきひろ氏に関する情報」はそこからたどれるリンク先などに記述されていると考えられる．

② タイプ(ii)において，表見出しと内容セルは，「50gまでの定形郵便は90円かかる」「50gまでの定形外郵便は120円かかる」…といった現象集合の総称概念「郵便物の料金」と，それに含まれる特定条件下での一つの具体的現象「120円 (かかる)」，という組合せであり，検索意図としては例えば，その現象が起こる条件を知りたい，というケースが想定される．その場合は，行/列見出しに知りたい情報が存在することになる．

③ (B)も(A)同様，要素間に強い関係性があり，検索状況が十分想定可能であることから，表構造における係り受けパターンとして抽出する方がよい．(B*)についても同様である．

(C) 行見出しと列見出し

① タイプ(i)において，行見出しと列見出しは，ある一つの実体/現象に対してのある属性，という関係にあり，その属性値が求められていると考えられる．例えば，具体的なスクリプト言語名と「開発者」というキーワードの組合せで検索が行われたときには「○○の開発者が知りたい」という検索意図が推定できる．この場合には，キーワードとして指定された見出しの行と列とが直交するセルにその属性値が存在する．

② タイプ(ii)において，行見出しと列見出しは，一方の条件「定形外郵便」をもう一方の条件「100gまで」で更に絞り込む関係であり，すなわちこの場合の検索意図は「定形外郵便かつ100gまでときにかかる料金が知りたい」と推定できる．この場合，両方の条件に当てはまる値「140円」は，①と同様に指定された行と列とが交差するセルに存在する．

③ ①②ともに，直交するセルに検索者の求めていた答が記載されていることが期待されることから，(C)を表構造における係り受けパターンとして抽出する．

(D) 行（または列）見出しと内容セル

① タイプ(i)において，行見出しと同行の内容セルは，ある実体/現象のキーと，その実体/現象の属性値，という関係にある．列見出しと同列の内容セルは，属性名と，ある実体/現象の属性値，という関係にある．これらはいずれも，密接な意味的関係があるといえる．実体/現象のキーと，その実体/現象の属性値の場合は，実体/現象の指定された属性値以外の属性値を求めている場合や，特定の実体や現象により限定される属性値に関する情報を求めている場合などが考えられる．

例えば「Ruby」と「まつもとゆきひろ」というキーワードの組合せの場合，「まつもとゆきひろ氏が開発したRuby」と解釈し，Rubyの発表年や実行速度などが求められているという場合と，「Rubyを開発したまつもとゆきひろ氏」と解釈し，まつもと氏に関する情報が求められているという場合が考えられる．前者の場合には，該当行に求めている情報が存在すると考えられる．後者の場合には当該の表には，求められている情報（一つの属性値に関する詳細情報）はない可能性が高いが，内容セル内でキーワードがリンクをもつ場合，そのリンク先に属性値に関する詳細情報が書かれていると考えることができる．

また属性名と属性値の組合せの場合，属性名と属性値から実体や現象のクラスが特定できるケース（「本社所在地　浜松市」「開戦日　12月8日」など）においては「その条件に該当する実体や現象を調べたい」といった状況が想定できる．この場合も指定された内容セルと同じ行に調べたい実体/現象のほかの属性値が書かれていると考えられるため，検索に有効であることが期待できる．

一方，行見出しと別行の内容セルは，ある属性名と，それとは別の属性の（ある実体/現象の）値，または，ある実体/現象のキーと，それとは別の実体/現象の属性値，という関係にある．例えば「Python」と「Wall」など，検索キーワードとしてあり得ない組合せではないが，ほとんど関係のないものも多分に含まれる可能性が大きい．

② タイプ(ii)において，行/列見出しと同行/同列の内容セルの場合は，条件の一つとその条件に該当する値の一つという関係である（例えば「定形外郵便」「120円」など）．その値を規定する別の条件が知りたい（「定形外郵便物で，120円で送れるのは何グラムまでか」）という状況が考えられる．この場合，指定された内容セルの行見出し，列見出しのうちで，キーワードとして指定されなかったものが求める情報である．

しかし，条件（見出し）と別の行/列である場合は，内容セルには見出しとは別の条件に基づく値が記されているのであり，意味的な関係は極めて薄いと考えられる．

③ 見出しと内容セルが同行または同列の場合は，検索に有効と思われる意味的な関係が想定できる．一方，見出しと内容セルが同行または同列にない場合は，意味的関係がない可能性が高いことから，(D) は同行/同列についてのみ，表構造における係り受けパターンとして抽出すればよいと思われる．

(E) 同一見出し内または同一内容セル内

①② タイプ (i)(ii) を問わず，同一見出し内，同一セル内の場合は，構文解析によって求められる関係がある．

③ そこで (E) は，該当箇所をとりあえず抽出しておき，文構造による係り受けパターンの解析結果に判断を委ねることとする．

(F) 行（または列）見出し同士（別々のセル）

① タイプ (i) における列見出しとは，その列に整理されている属性名であり，列見出し同士は，それぞれ別の属性概念同士ということになり，直接的な関係はもたない．行見出しは，1行で表現される実体/現象のキー概念であり，行見出し同士は並列の関係となる．

② タイプ (ii) において，行見出し同士/列見出し同士とは，並列の関係であり，同時には成立しない条件同士である．

③ これらはいずれも，検索の状況を想定しにくく，表構造の係り受けパターンとして採択すべきではないと思われる．

なお，キーの属性見出し（1行1列目）は表見出しであるケースが多いため，行/列見出しには含めず，表見出しとみなすものとする．

(G) 異なる内容セル

① タイプ (i) においては，別々の内容セルでも，同行の場合は同じ実体/現象についての情報であり，二つの属性値からそれらを属性値としてもつ実体や現象を調べたい場合や，一方の属性値をもつ特定の実体/現象を指定することによって限定される他方の属性値に関する情報を調べたい場合などが考えられる．前者の場合，開発者名と特徴から該当するスクリプト言語を調べたい場合などであり，この場合，行見出しに答がある．後者の場合は，特定の言語を指定するためにスクリプト言語の一つの特徴を指定し，その言語の開発者であるという限定を加えて，もう一つのキーワードの人名を指定し，その人に関する情報を求める場合などであり，この場合，人名からリンクが張られていればそこに求むる情報があると考えられる．ただし同列の場合は，並列の関係であり（別々の実体/現象の属性値を差す），その場合の2者の関係は薄いことが多い．

同行/同列の場合以外は，別の実体/現象の属性値同士であり，関連が薄い．

② 一方，タイプ (ii) において内容セルは，同行・同列かどうかにかかわらずすべて均一な概念で構成され，これに該当する検索の状況を想定しにくい．よって2者の関係は薄いと考えるのが妥当である．

③ タイプ (i) の同行の場合については，検索の状況が想定でき，求める情報も表の中にあると考えられることから検索に有効であることが期待できる．行が基本単位となるタイプ (i) の表で同列の場合には関係は薄いが，列が基本単位となるケースも考慮する必要がある．よって，タイプ (i) については，同行/同列についてのみ，表構造における係り受けパターンとして抽出すれば十分と思われる．

(H) ページタイトルと表見出し

①② ほとんどのWebページにはページタイトルが付けられているが，そのスコープ内に表は含まれるものであり，表にとってページタイトルは無視できない存在である．ページタイトルを大見出し，表見出しを小見出しととらえることも可能である．すなわち，何らかの意味的関係にある可能性が高い．例えば「マレーシア　祝日」というキーワードで検索するとき，ページ全体がマレーシアに関する情報となっており，その中に祝日一覧表が含まれていれば有用である．このページタイトルに「マレーシア」が含まれていて，祝日一覧表には「祝日」というタイトルが付けられている状況は，自然である．

③ 検索に有効な事例が存在し得ることから，係り受けパターンとして抽出すべきである．

以上の表の性質から，(F) は表構造における係り受けパターンとはいえないため，取り扱うべき組合せから除く．また (D) は同行/列の見出しと内容セルのみを対象とすればよい．(G) は，タイプ (i) では同行/列のみを対象とすべきであるのに対しタイプ (ii) では除くべきと，タイプ (i) とタイプ (ii) とで違いがあった．今回検討した限りにおいては唯一の相違だったため，タイプ (i)(ii) を別々に処理することは省略し，(G) は同行/列の内容セル同士の場合のみを対象とすることにする．よって，以下のパターンを拾い上げればよいと整理できる．

(A*) ページタイトル+行（列）見出し

(B*) ページタイトル+内容セル

表 1　表構造における係り受けパターン
Table 1　Dependency patterns in table structure.

記号	検索キーワードの出現パターン
(A*)	ページタイトル+行（列）見出し
(B*)①	ページタイトル+内容セル（リンク）
(B*)（①除く）	ページタイトル+内容セル（①を除く）
(A)	表見出し+行（列）見出し
(B)①	表見出し+内容セル（リンク）
(B)（①除く）	表見出し+内容セル（①を除く）
(C)	行見出し+列見出し
(D')①	行（列）見出し+同行（列）の内容セル（リンク）
(D')（①除く）	行（列）見出し+同行（列）の内容セル（①を除く）
(E)	同一セル（見出し，内容セル，ページタイトル問わず）
(G')①	同行（列）の内容セル同士（リンク）
(G')（①除く）	同行（列）の内容セル同士（①を除く）
(H)	ページタイトル+表見出し

(A)　表見出し+行（列）見出し
(B)　表見出し+内容セル
(C)　行見出し+列見出し
(D')　行（列）見出し+同行（列）の内容セル
(E)　同一セル（見出し，内容セル問わず）
(G')　同行（列）の異なる内容セル
(H)　ページタイトルと表見出し

また更に，HTML 文書の特徴として，リンクの存在を考慮すべきである．内容セルに存在するキーワードがリンクとなっている場合には，リンクとなっているところが内容セルの主たる内容であり，リンクの先にその詳細な情報が置かれている可能性も高いと考えられる．そこで，内容セルに関して（(B) と (D') と (G')）は，①キーワードがリンクになっている場合と，そうではない場合，それぞれ区別して分析を行うこととする．

以上の議論を整理し，HTML 文書内の表構造における 2 語が出現するパターンについて，何らかの意味的関係にあると思われるパターンを中心に整理を行った（表 1）．

なお，表の中に存在する検索キーワードの位置について厳密に議論するには，セルの中の位置について，分けて議論を行う必要がある．「セルの中のキーワードの位置」とは，セルを構成する文において，キーワードがどのような位置に存在しているか，ということである．例えば，文のヘッドに出現しているケースと，あるいは従属文の中に出現しているケースとでは，明らかにキーワードの重要度が異なってくる．しかしながら今回は，セルの中の位置は考慮せず，(近似として) セルの中のキーワードの有無だけで議論を行う．

表 2　ウェブページ中の表を目視により分析調査した結果
Table 2　Result of research by viewing tables in web pages.

n=500（ページ）

表の係り受けパターン	適合	全体	適合ページの割合
(A*)	22	23	95.7%
(B*)①	13	17	76.5%
(B*)②	2	2	100.0%
(B*)③	2	2	100.0%
(B*)（①②③除く）	3	8	37.5%
(A)	13	14	92.9%
(B)①	9	11	81.8%
(B)②	1	1	100.0%
(B)③	2	3	66.7%
(B)（①②③除く）	5	10	50.0%
(C)	20	20	100.0%
(D')①	2	2	100.0%
(D')（①除く）	1	5	20.0%
(E)	26	28	92.9%
(G')①	3	7	42.9%
(G')（①除く）	0	4	0.0%
(H)	11	13	84.6%
いずれにも該当せず	101	330	30.6%
計	236	500	47.2%

3.2　適合可能性に基づいた係り受けパターンの検討

前節では，二つのキーワードがともに表内の要素となる場合と，一方がページタイトルにあり，他方が表内の要素にある場合とについて検討し，抽出する意味をもち得るものとして表 1 のパターンを挙げた．ここではそれらのパターンについて，現実的な有用性を検証するため，2 語で検索した結果ページ（上位 100 件×6 組）(注3) の中からページ中にテーブルタグが存在する 500 ページを対象として目視による分析調査を行った（表 2）．分析中，(B)（①除く）と (B*)（①除く）については，以下に該当する場合に適合ページである可能性が高いという傾向が見て取れたため，細分化を行った．

②　キーワード自体はリンクではないが，同行にリンクが存在しているケース(注4)

(注3)：キーワードの選定には，与えられた検索課題（「どこのメーカーのノートパソコンが一番売れているのか？」など）に従ってキーワードと検索意図を記入してもらうアンケートを利用した．その中から客観的に適合判定基準（「ノート PC のシェアや人気ランキング等が紹介されていれば適合」など）が作れるものを選び，その判定基準に従って適合/不適合判断を行った．
(注4)：リンク集のページでよく見受けられる．(キーワードを含まない) サイト名がリンクになっており，その横のサイト紹介文の中にキーワードが存在しているようなケースである．

③ （①②いずれにも該当しないながら）キーワードが一つの表内に頻出（暫定的にしきい値を5回以上とする）するケース

すると (B*)（①②③除く），(D')（①除く），(G')①，(G')（①除く）の4パターン以外に関して，検索者が求める情報が，表の中またはリンク先に存在している可能性が非常に高い傾向にあることが分かった（500ページ中，適合ページは236ページ（47.2%）であったのに対し，この14パターンのいずれかに該当するページは146ページあり，うち適合ページは128ページ（87.7%）であった）．

(B*)（①②③除く）が外れる原因としては，内容セルに存在しているキーワードに関する情報を求めているが，リンクもなく，頻出でもないことから，情報が足りなかったため不適合となった可能性と，表の見出しに比べ，ページタイトルは内容セルとのかかわりが弱まることに起因する可能性が挙げられる．(D')（①除く）については，タイプ(i)の表において，実体/現象の指定された属性値以外の属性値を求めている場合と，特定の実体や現象により限定される属性値に関する情報を求めている場合とを想定したが，実際には後者の場合が多く，その場合リンク先で属性値に関する詳細情報が説明されていない場合には不適合と判定されたためと考えられる．また，(G') については，一つのレコードが一つの名詞句を構成できる場合や一つの動詞に係る場合は2語にある程度の関連が認められるが，必ずしもそうではなく，例えば，一つのレコードでもセルによって別の述語をもつ場合もあり，このようなケースでは同じ行のセル同士でも密接な関係にあるとはいいがたい．このため，想定したような検索意図で2語が用いられることが，現実には少なく，(G') に該当する適合ページが少なかったと考えられる．

実験結果と以上の理由を踏まえ，(B*)（①②③除く）と (D')（①除く），(G')，更に文の係り受け判定に依存する (E) を除いたパターンを，表の係り受け判定として採用することにする．

4. 表の抽出

4.1 HTML文書中に存在する表の判別処理

表の係り受け判定を実際にプログラムで適用するためには，「表」を正しく抽出できるということが前提となる．Wangら[12]が指摘するように，ウェブページに出現する多くのテーブルタグは，テーブルを表すために用いられるのではなく，レイアウトを制御することにも用いられる．彼らは，テーブルタグを用いて構成されたテーブルの中から「本物の表」を抽出するため，レイアウト構造（タグ）やセル内の字種などの特徴と，従来のテキスト分類の手法を組み合わせ，高い精度（精度：97.5%，再現率：94.3%）で判別を実現している．

確かに，レイアウトのために用いられるテーブルは，同じ行や列のセル同士に意味的な深い関係をもたないどころか，全く別のテーマを扱うことさえある．したがって，我々にとっても，テーブルタグを用いて書かれているテーブルを「意味的に表を表すテーブル（タイプA)」と「レイアウトを制御することだけに用いられるテーブル（タイプB)」とに弁別することは，正しく係りを判定する上で必要不可欠である．

そこで我々も，独自のアルゴリズムでテーブルタグをタイプAとタイプBに分類し，タイプBについては表の係り受け判定処理の対象外とするための判別プログラムの開発を行い，表判定の前処理として組み込んだ．

判定戦略として，

- 句点の存在の有無
- タグ以外のテキストの存在の有無
- 色の指定の有無
- 線の幅指定の有無
- セルの数
- リンクのあるセルの数
- 文字のあるセルの数

を独立変数として用い，目視で判定したデータ（対象：120ページ，1909テーブル）を従属変数（教師データ）としてC4.5[(注5)]によって最適な決定木を作成し，それをもとにできるだけタイプAを取りこぼさないように改良を行った．オープンテスト（対象：60ページ，657テーブル）の結果は，表3のとおりである．目視とプログラムの判定が一致しなかったのは5.8%+4.7%=10.5%で，Wangらの精度には及ばないもののおよそ9割は正しく判定されており，我々は実質上使用できるレベルに達していると考えている．

4.2 表の見出しの抽出処理

表の見出しをプログラムで抽出する必要がある．以下の条件に基づいた見出しの抽出処理を行う．

- 表見出し：CAPTIONタグ，(タイプAのテーブルの）1行1列目セル，見出し用テーブル，親テー

(注5)：http://www.cse.unsw.edu.au/~quinlan/

表 3 表（タイプ A）/レイアウト（タイプ B）判定結果
Table 3 Result of judgment of table type.

n=675（テーブル）

		プログラム	
		タイプ A	タイプ B
目視	タイプ A	23.6%	5.8%
	判定困難	0.3%	0.8%
	タイプ B	4.7%	64.8%

ブル

- 行見出し：（タイプ A のテーブルの）2 行目以降の 1 列目
- 列見出し：（タイプ A のテーブルの）2 列目以降の 1 行目
- 以上の条件に該当していても，その中に句点が含まれる場合は，見出しではないとみなす

ここで，見出し用テーブルとは，タイプ A の表見出しとしての役割しかもたないタイプ B のテーブルであり，タイプ B のテーブル（ただしタグを除く文字が含まれるセルが二つ以内のもの）とタイプ A のテーブルが続けて現れるときの前者のテーブルを指す．親テーブルとは，セルの中にタイプ A のテーブルを含むタイプ B のテーブルであり，タイプ B の中でタイプ A のテーブルを内部に含むテーブルを指す．なお，これらの見出し判定アルゴリズムはヒューリスティックに基づくものである．

なお，表によっては「1 行目」「1 列目」が見出しとなっていないケース（行見出しが存在しない，複数行/列の見出し，多段組の表など）や，見出しが階層構造をもつ場合もあり得る．このような複雑な表の扱いについては，田仲ら [13] や大西ら [14] による研究を参考に，別途検討していきたい．

5. システム構成

本システム（図 4）は，既存の検索エンジンを利用したフィルタリングツールである．

現在のところ，実験には既存の検索エンジンとして Google が提供する API(注6)，プログラミング言語として Ruby を用いており，システム全体については Vine Linux 上で動作している．

5.1 システム全体の流れ

（1） まず，ユーザによってウェブブラウザから検索キーワードが入力されると，既存の検索エンジンにキーワードを渡し，検索結果のウェブページを取得する．

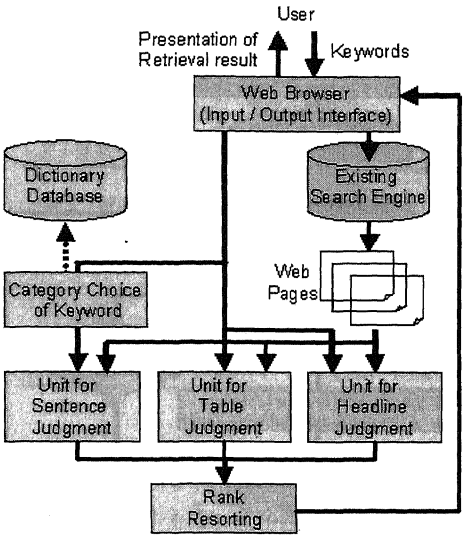

図 4 システム全体図
Fig. 4 System configuration.

（2） 各ウェブページは，今回提案した表構造による係り受け判定ユニットのほか，別途開発した文判定ユニット [1] 及び見出し判定ユニット [2] により，係り受けパターンにマッチするか否かを判定される．

- 表判定ユニットについては，**5.2** で詳述する．
- 文判定ユニットでは，各キーワードをパーザの概念階層辞書を用いて実体，現象，属性，値という四つのカテゴリーのいずれかに分ける．そのキーワードカテゴリーを用いて係り受けパターンの候補を挙げる．次いで，キーワードを含む文がそのうちのどれかのパターンに該当するか否かを調べる．
- 見出し判定ユニットでは，二つのキーワードがそれぞれ見出しの一部として存在し，かつその見出し間に親子関係があるか否かを調べる．

（3） 各判定ユニットで，パターンにマッチしたページをキーワード間の意味的関係が強いとみなし，一定のスコア（「一つ以上のパターンにマッチしたら 1，一つもマッチしなかったら 0」）を与える．複数のユニットを組み合わせる場合にはこのスコアを加算する．このスコアを第 1 ソートキーとして降順に並べ，第 2 ソートキーをもとの検索エンジンの順位とする手法でランキングを行う．

（4） こうして並べ換えられた検索結果はユーザに

(注6)：Google SOAP Search API http://www.google.com/apis/

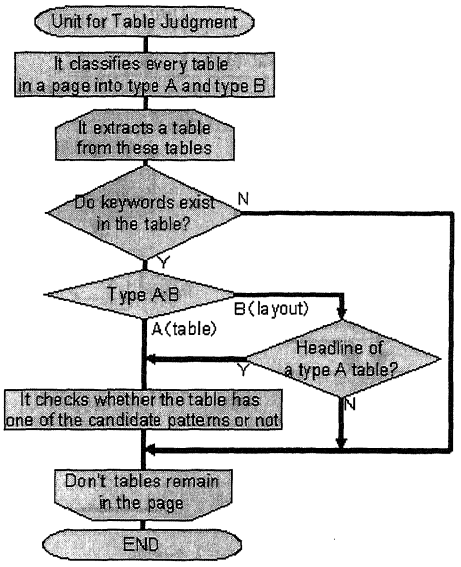

図5 表判定ユニット
Fig. 5 Unit for the table judgment.

提示される.

5.2 表判定ユニット

表判定ユニットの処理の流れを図5に示す.

(1) まず, ページ中に存在するすべてのテーブルを抽出する.

(2) 抽出したすべてのテーブルについて, あらかじめタイプAかタイプBかを判定しておく.

(3) テーブルを出現順に取り出し, 中に検索キーワードが存在するか否かをチェックする.

(4) キーワードが存在するテーブルに対し, キーワードの位置の解析 (何行何列目か, 見出しか内容セルかといった判断) を行う.

(5) そして表1のパターン ((B*) (①②③除く), (D') (①除く), (G') を除く) に該当するか否かを調べる.

(6) 未処理のテーブルがなくなるまで, (3)(4)(5) を繰り返す.

6. 評価実験

本章では, 表判定の効果, 有効性を検討する. そのためまず, データセットを作成し, 文判定と見出し判定のみを用いる場合に較べ表判定を加えることで検索性能に向上が見られるか否かを評価した. データセットの作成方法及び評価手法については **6.1** で述べる. 結果の詳細は **6.2** で述べるが, 文判定と見出し判定で得られた精度をほぼ維持したまま高い再現率で候補を抽出できることが示せた. しかし同時に検索キーワードによって効果に差があることが分かった.

比較的有効であるキーワードを確認したところ, 検索者が求める情報が表構造で整理されていることが予想できるようなキーワードであることが多いという傾向が見てとれた. すなわち, 該当する情報が表構造となって表現されやすいキーワードと, されにくいキーワードがあり, 表判定は前者 (以下,「表をイメージさせるキーワード」という) に特に有効であると考えられる. 再現率の大幅な向上は, そのようなキーワードに対して文判定と見出し判定では抽出できない候補を拾い上げることが可能になったことによるもの考えられる.

一方, 効果の薄いキーワードの場合, 表判定を加えることで精度の低下を招くケースもあった. 実験結果の平均の精度は文判定と表判定のみの場合と表判定を加えた場合とはほぼ同等だが, 効果の薄いキーワードで検索を行う場合についていうと, むしろ表判定は使用しない方がよいということになる. そこで, 表判定を利用するか否かの選択をオプションとして提供し, ユーザの判断に従って表判定を適用するという活用方法を想定し,

(1) キーワードから, それが「表をイメージさせるもの」であるか否かという判断が, 個人差なく一般的に行えるか否か

(2)「表をイメージさせるキーワード」とされたものの範囲内で, 表判定がどの程度有効に機能するかという2点を確認することとした.

その結果,「表をイメージさせるキーワードである」という判断は比較的個人によらない安定したものであり,「表をイメージさせるもの」とされた範囲内で表判定を利用すると大きな効果を引き出せることが分かった. この実験とその具体的な結果については **6.3** で述べる. **6.4** では, 表判定を導入することにより精度が下がる要因, 文判定・見出し判定・表判定の三つを用いた総合的判定において現段階で対処できていない問題点について考察する.

6.1 評価手法

6.1.1 評価データの作成

オープンテストにより本手法を評価するため, データセットを作成する. まず, 以下の (1)～(3) の手順で検索キーワード対の設定を試み, 118組を選んだ.

(1) ウェブページを10万ページ取得して形態素解析を行う.

（2） 一般的すぎる語では漠然としていて検索キーワードとしてふさわしくないため，TF-IDF で上位となった語から，人手で検索キーワードとして妥当な組合せを 2000 組つくる．ただし，語を選ぶ際，共起確率の高い語同士を複合語とすることも認める．

（3） （2）の作業にかかわらない複数人（今回は 3 人）が検索意図を推定し，明文化する．例えば，以下は「ウイルス　対策」というキーワードの場合であるが，このように 3 人が推定した意図が一致したものを選ぶ．

A 氏「ウイルス対策方法を紹介しているサイト」
B 氏「ウイルスの対策方法が書いてあるサイト」
C 氏「ウイルス対策の方法を知りたい」

この意図に沿って（不適合の判断も補足して）下記のように適否の判定基準を明文化する．

「ウイルスの対策方法が記載されていれば適合，ウイルス対策ベンダーやウイルス対策機能付き商品についてのニュース記事は不適合」

この方法で適否の判定基準を決定できた検索キーワード対 118 組について Google を用いて検索を行い，上位 100 件までにランキングされたページをダウンロードし，ダウンロードに失敗したページ及びバイナリーファイルを除いた 11,776 ページについて，適合/不適合の判定を行った．これを評価用のデータとする．判定は，上述の明文化された判定基準に基づいて 3 名の学生の合議により行った．なお，リンク先を実際に確認しなくても（するまでもなく）検索意図にかなった情報がリンク先に存在していることが明らかな場合はこれも含むものとする．

6.1.2 評価尺度

比較の評価指標には，各キーワード対当り約 100 件づつの評価データ内でシステムに適合と判定された集合に対する精度・リランキング後の上位 20 件中の適合ページ数・MAP を用いる．精度は，**5.1** で述べたスコアが 1 以上のページを適合として算出する．一方適合ページ数と MAP は，**5.1** のスコアリングに基づいて順位の並べ換えを行い算出する．MAP（Mean Average Precision）とは，検索課題ごとの平均精度の平均であり，ランキングの善しあしを評価するための指標である．R を適合文書の総数，n を出力文書数とし，

$$z_i = \begin{cases} 1 & (\text{順位 } i \text{ 位の文書が適合}) \\ 0 & (\text{順位 } i \text{ 位の文書が不適合}) \end{cases}$$

表 4　提案手法の精度・再現率・F 値
Table 4　Precision, recall and F-value of our method.

n=11,776（ページ数）	精度	再現率	F 値
文＋見出し判定	65.5%	58.8%	62.0%
文＋見出し＋表判定	63.6%	71.2%	67.1%

表 5　適合ページ数と MAP（全データ対象）
Table 5　The number of relevant pages and MAP.

	適合ページ数	MAP
(I) もとのランキング	1428	0.601
(II) 表判定のみ	1548	0.639
(III) 文＋見出し判定	1590	0.666
(IV) 文＋見出し＋表判定	1598	0.667

とする．このとき，平均精度 v は，次式で求められる．

$$v = \frac{1}{R}\sum_{i=1}^{n}\frac{z_i}{i}\left(1+\sum_{k=1}^{i-1}z_k\right) \tag{1}$$

MAP のもとになる平均精度は，上位に存在する適合文書の方が下位より重視される指標であり，20 位中の適合ページ数は同じであっても，上位にランキングされている適合ページの量によって平均精度は異なることになる．なお，手法の性能比較を行う場合，MAP に 0.05 程度の差があれば有意差があるといわれているようであるが，より厳密には平均値の差の検定を実行すべきである [15]．そこで有意水準 α =0.05 で t-検定を行う．

6.2 全検索課題を対象とした総合的判定の評価

提案手法を適用した場合の検索性能を確認するため，**6.1.1** で述べたデータセット全 118 組の各検索結果上位 100 件に対して，文判定，見出し判定のみを用いる場合と，それらに加えて表判定も併用する場合について比較を行った．まず，それぞれの判定基準に基づき適合と判断された全ページに対する精度・再現率・F 値を算出し，次いで各課題についてランキングを行った結果上位 20 位までの適合ページ数及び 100 位までの MAP を算出した．

表 4 に示すとおり，文判定と見出し判定のみの場合と比較し，表判定を含めることによって，精度がわずかに落ちるものの再現率を大幅に向上させることができた．一方で，上位 20 件中の適合ページ数と MAP（表 5）では，表判定を加えることによる差はほとんど見受けられなかった．

差が生じなかった原因は，前述したように，表判定が精度向上に有効であるキーワードと逆効果となってしまうキーワードが存在し，全体的には相殺されてし

まったことによる．

6.3 表をイメージさせるキーワードを対象とした評価

まず，「(1) キーワードから，それが「表をイメージさせるもの」であるか否かという判断が個人差なく一般的に行えるか否か」を確認するために，被験者4名に，**6.1.1** で述べた118組それぞれの検索キーワードから検索者が欲しいであろう情報がウェブページ中にどのように存在していればベストかを想像させ，(a)〜(c) のいずれに該当するかを判別させた．

(a) 個別の実体または現象及びそれらがもつ属性や値ではなく，それらが複数集まった集合としての情報が欲しいとき，それが一つの表に集積されているイメージ

(b) 個別の実体または現象の属性値について知りたいとき，表の中のある1箇所に答が記述されているイメージ

(c) 表にはなっていない

被験者には，以下のような例を示して説明し，イメージを具体化してもらった．

(a)「浜松　ラーメン」→ 浜松のおいしいラーメン屋さん一覧が見たい．

(b)「マレーシア　首都」→ マレーシアに関する表があって，その中の一つのセルに答があるはず．

(c)「Linux　インストール」→ 方法が知りたいので表ではない．

その結果，表をイメージさせる/させないで4名の意見が半々に割れたものが11.9%，1名のみ異なったものが42.4%，4名とも揃ったものは45.7%であった．このことから，ユーザの判断はある程度安定して一致すると考えられる．

次に，「(2) 表をイメージさせるキーワード」とされたものの範囲内で，表判定がどの程度有効に機能するか」を確認するために，4名中3名以上が (a) または (b) と答えたキーワード対から無作為に20組（表6）を抽出し，それらに対する提案方法の性能を調べた．

この20組の集計結果は，表7に示すとおりである．上位20件における適合ページの数が，もとの順位 (I) と比較して平均1.5ページ，増えた計算になる．また (III) の文判定と見出し判定によるランキングと比べても，(IV) の表判定を加えた結果の方が，平均して1.0ページ多いという結果になった．なお，検索課題 (2)(4) は表判定を加えたことにより適合ページ数が少なくなっているが，これらの場合でも「上位10位以内の適合ページ数」で比較すると表判定を除いた場合より表判定を加えた場合の方が多いことを確認できている(注7)．更に，精度・MAP とも (IV) が最も良い．そこで平均精度について t-検定を実施した．その結果，(I) と (II)，(III) と (IV) については差は認められるものの，有意な差とは認められなかった．しかし，(I) と (III) の比較において有意差が得られなかったのに対し，(I) と (IV) では有意な差 ($p = 0.029 < 0.05$) が生じた．

表6 抽出したキーワード対
Table 6 Extracted keyword pairs.

	検索キーワード
(1)	欧州　サッカーチーム
(2)	中国　映画監督
(3)	米国　ロックグループ
(4)	アジア　アイドル
(5)	日本代表　野球選手
(6)	女性　政治家
(7)	料理　学校
(8)	ペット　用品店
(9)	考古学　研究所
(10)	神奈川　劇場
(11)	長野　映画館
(12)	京都　庭園
(13)	ドメイン　登録料
(14)	北京オリンピック　開催日
(15)	コミック　発売カレンダー
(16)	ビジネス　書籍
(17)	ディズニー　映画
(18)	CD　ランキング
(19)	お笑い　ライブ
(20)	テーマパーク　比較

これらの結果より，全体の再現率よりも上位の候補の精度を重視する場合においても，表判定を含めた手法が上位の適合ページ数を増やすことができるといえる．かつ，前節で示したように，表判定を含めた手法 (IV) は，文+見出し判定のみの手法 (III) に較べ，再現率の点では大きく改善できるので，総合的にフィルタリングツールとして (IV) が有効であると考えられる．

6.4 考　　察

6.4.1 表判定によって精度が下がる要因

今回，文判定・見出し判定に表判定を加えたことで精度が下がった事例について，原因の分析を行った．118組 11,776 ページにおいて，表判定のみで適合と判断された不適合ページの中から無作為に280ページ (20%) を抽出し，調査した結果を表8に示す．なお，このうち要因 (2)(3)(5) は，文判定・見出し判

(注7)：(2) は9ページが10ページに，(4) は7ページが8ページに適合ページが増加した．

表7 表をイメージするキーワード対の上位適合ページ数と精度，MAP
Table 7 The number of relevant pages, precision and MAP.

n=2000（ページ数）

	(I)	(II)	(III)	(IV)
(1)	13	13	12	13
(2)	15	17	17	16
(3)	9	9	9	9
(4)	13	18	15	13
(5)	4	6	4	7
(6)	6	4	4	5
(7)	13	16	14	15
(8)	15	11	15	16
(9)	13	18	17	17
(10)	10	18	15	18
(11)	19	18	15	18
(12)	19	16	17	18
(13)	16	17	15	17
(14)	2	2	2	2
(15)	7	6	8	9
(16)	12	14	13	14
(17)	11	9	12	13
(18)	16	18	19	19
(19)	8	12	11	13
(20)	5	3	2	4
適合ページ数計	226	245	236	256
MAP	0.565	0.598	0.609	0.617
精度	0.450	0.652	0.580	0.586

(I)…もとのランキング
(II)…表判定のみによるランキング
(III)…文＋見出し判定によるランキング
(IV)…文＋見出し＋表判定によるランキング

定を含めた総合的判定においても共通に見られる問題点であり，表判定固有の問題ではないため，次項で考察する．

全体の半分弱を占める要因(1)とは，2語がページ中に存在し，表1のパターンに該当するが，実際はそれぞれ別の語と修飾–被修飾関係にあり，別の内容を表すような場合である．例えば「『旅行内容・料金の比較』サイトにおいて表の中に『テーマパーク』が含まれているページ」を『テーマパーク　比較』で検索した場合の適合ページと判断してしまうようなケースである．これは，現在のところキーワードが表中のどことどこにあるかという位置関係しか見ていないことに起因する．本来は，品詞や意味からどういう修飾–被修飾関係を構成するかを推定して，文構造で表されている修飾–被修飾関係と矛盾しないか等をチェックする必要がある．例えば上の例では，テーマパークは連体助詞を介してサ変名詞「比較」に係る，あるいは対象格補語として「比較する」に係ると推定される．ところが，表中の「比較」には既に連体助詞を介して「旅行内容・料金」が係っているため，この場合，「テーマパーク」と「比較」の係りは認めるべきではない．

表8 精度低下の要因
Table 8 The causes of low precision.

n=280（ページ数）

	要因	割合
(1)	2語が意味的に係っていない	44.3%
(2)	ページタイプが検索意図にそぐわない	28.2%
(3)	語が検索意図とは異なる意味で使われている	10.4%
(4)	2語が更に別の語に係って意図から外れる	7.1%
(5)	その他	10.0%

また，例えば「テーマパーク周辺ホテル」と「比較」を表中で発見して「テーマパーク」と「比較」の間に意味的関係があると判定してしまうような事例（「ホテルの比較」であり「テーマパークの比較」ではない）も要因(1)に分類されている．これに対応するためには，**3.1**で述べた「セルの中のキーワードの位置」を考慮する必要がある．

要因(4)は，例えば「ディズニー」と「映画」という検索キーワードにおいて，「ディズニー」と「映画で使われた楽曲の紹介」を表中で発見して意味的関係があると判定した場合である．この場合，「ディズニー」と「映画」の間には確かに意味的関係を認めることはできるが，表内に表現されている情報は，「映画」に関するものではなく，「楽曲」に関するものであると考えられ，検索意図には合致しないことが多い．このようなケースに対しても，「セルの中のキーワードの位置」を考慮することが有効であると考えられる．

6.4.2 総合的判定における精度低下要因

表判定・文判定・見出し判定を併せた総合的判定における，現段階で対処できていない精度低下の要因について述べる．

表8の要因(2)は，ウェブページの多様性に起因したものである．ウェブ空間には，論文など学術的なものから，企業や商品の宣伝，個人の日記や掲示板などまでが区別なく混在している．これでは，ページ内の検索キーワードが意味的関係をもって出現していたとしても，検索意図にそぐわないことがあり得る．例えば専門用語の意味が知りたいときに書籍紹介のページやシラバスが与えられても検索者は満足できない．

この要因(2)に該当するページタイプには，大きく分けて以下の三つのタイプが確認された．

- 個人的な内容・・・ブログ，日記，体験談（約2割）
- 過去の内容・・・ニュース，キャンペーン・新製品

などのお知らせ，過去のオークション・プレゼント，求人（約5割）
- 見出しのみで中身（解説）がない・・・書籍・授業・セミナー紹介（約3割）

これらについては，文末表現（助動詞など）に着目することによる対応策を，現在検討中である[16].

要因(3)は多義語の問題である．例えば「自動車保険　ポイント」というキーワードを利用し，自動車保険を契約する上で参考になる解説ページを検索するケースにおいて，「自動車保険でポイントを貯める」という内容のページを不適合と判定することは現状不可能であり，解決は非常に困難な問題である．

要因(5)は，要因(2)の中の「見出しのみで中身（解説）がない」とも関連するが，パターンにマッチした箇所が，ページの中での扱い（重要度）が小さく（属性値などの）知りたい詳細な情報が書かれていない，といったケースが含まれる．これを判断するためには，キーワードが存在している構造自身が，その支配範囲にどれだけの情報量をもっているかという情報が必要である．支配範囲がほとんど（もちろんリンクも）ない場合，少なくとも，そのキーワードの解説はページ内に存在しない可能性は高いと思われる．

6.4.3 総合的判定における再現率低下要因

総合的判定において適合ページを取りこぼす主な要因を調査したところ，以下に挙げるケースに該当する事例が多く見受けられた．

まず，キーワードの1語が見出しに含まれ，もう1語がその見出しの支配範囲下にある地の文に含まれる場合である．現在の見出し判定の戦略では，見出し間の親子関係を対象とするため，片方の語が地の文にしか含まれない場合は対象とならない．一方のキーワードの支配範囲下において任意の場所にもう一方のキーワードが出現さえしていればよい，としてしまうと多くの不適合ページを誤って適合とみなしてしまうことになるが，地の文でも
- リンク
- 頻出
- 1行目
- カギ括弧による強調

などと条件を絞り込むことによって，精度を下げずに再現率を上げられる見込みはあると思われる．

二つ目は，一方のキーワードの同義語・類義語・下位語がもう一方のキーワードと意味的関係にある場合である．例えば「CD　ランキング」の検索キーワードに対し「シングル」と「ランキング」は係り受けパターンに該当する位置関係にある，などといったケースが多数見受けられた．

6.4.4 今後の課題

現在は表の係り受けパターンのうちのいずれかに該当するか否かといった判断しか行ってないが，今後，ユーザの判断あるいはキーワードの種類（品詞や意味）によって，利用するパターンを限定したり，個々のパターンに重みを付けてスコアリングに反映したりする可能性がある．

例えば，文の係り受け判定で得られた精度（precision）を悪化させないことを条件とするなら現在の基準が適当であるが，目的によってはもう少し緩やかな条件で，精度は多少下がっても再現率向上を優先する立場もあり得る．

また，文構造・見出し構造と合わせたスコアリング戦略全体についても，今後詳細に議論する余地がある．

7. むすび

本論文では，検索キーワード間の意味的関係を利用して適合ページをフィルタリングする手法について議論を行った．

意味的関係を表す構造には，文，表，見出しなど，様々な構造があり得るが，本論文では，表構造に焦点を当て，その構造の中に現れた意味的関係に基づき適合ページを判定する手法を提案した．更にそれを評価するために，既存の検索エンジンのフィルタリングツールを構築し，自作の評価用データセットを用いて実験を行い，システムの評価を行った．

評価においては，今回提案した表構造の係り受けパターンを拾い上げる手法を追加することによって，これまでの文構造・見出し構造による判定と比較し，精度をほぼ下げることなく再現率を10%以上向上させることを示した．また表をイメージさせるキーワードに限定すると，表判定を加えた総合判定が表判定を加えない場合と比べ上位20件中の適合ページ数，精度，MAPのいずれにおいても改善されることを確認した．

今後は**6.4**で示した本戦略で不十分である点を見直し，更なる精度及び再現率の向上を目指したいと考えている．

文　献

[1] 松本章代，小西達裕，高木　朗，小山照夫，三宅芳雄，伊東幸宏，“検索キーワード間の修飾–被修飾関係の詳細な分析に基づくwww検索性能の向上，”情処学論，vol.48,

no.10, pp.3386–3404, 2007.
[2] 西口直樹，松本章代，小西達裕，高木　朗，小山照夫，三宅芳雄，伊東幸宏，“見出しの階層関係を利用した www 検索精度の改善，” 信学技報，NLC 2005-114, 2006.
[3] B.J. Jansen, A. Spink, and T. Saracevic, “Real life, real users, and real needs: A study and analysis of user queries on the web,” Inf. Process. Manage., vol.36, pp.207–227, 2000.
[4] 風間一洋，原田昌紀，“Web サーチエンジン技術の高度化，” 人工知能誌，vol.16, no.4, pp.503–508, 2001.
[5] R. Zanibbi, D. Blostein, and J. Cordy, “A survey of table recognition: Models, observations, transformations and inferences,” Int. J. Document Analysis and Recognition, vol.7, no.1, pp.1–16, 2004.
[6] 大谷貴志，獅々堀正幹，柘植　覚，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，” 情処学自然言語処理研報，2002-NL-154, pp.137–144, 2003.
[7] 大前信弘，黄瀬浩一，“Web の表を対象とした属性の自動識別，” 情処学自然言語処理研報，2006-NL-171, pp.43–48, 2006.
[8] 吉田　稔，鳥澤健太郎，辻井潤一，“表形式からの情報抽出手法，” 言語処理学会第 6 回年次大会，pp.252–255, 2000.
[9] 板井久美，高須淳宏，安達　淳，“Html からの情報抽出と統合，” NII Journal, no.6, pp.9–19, 2003.
[10] 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，” 情処学自然言語処理研報，2002-NL-153, pp.65–72, 2003.
[11] 岩口義広，鄭　眠洙，獅々堀正幹，青江順一，“Www 空間上に存在する表構造の一索引化手法，” 情処学自然言語処理研報，2001-NL-142, pp.159–166, 2001.
[12] Y. Wang and J. Hu, “Detecting tables in html documents,” LNCS, vol.2423, pp.249–260, Springer-Verlag, 2002.
[13] 田仲正弘，石田　亨，“表構造の一般化に基づくオントロジの獲得，” 情処学論，vol.47, no.5, pp.1530–1537, 2006.
[14] 大西香織，田島敬史，“Web 上の表データの論理構造の発見，” Proc. Data Engineering Workshop, 2006.
[15] 岸田和明，岩山　真，江口浩二，“検索実験の方法と実際：ntcir ワークショップでの試み，” Pre-meeting Lecture at the NTCIR-3 Workshop, 2002.
[16] 包　直也，松本章代，鈴木雅人，“文末の表現に着目した閲覧者が受ける印象による web 文書のクラスタリング，” 情報処理学会第 69 回全国大会講演論文集，vol.2, pp.559–560, 2007.

（平成 19 年 5 月 29 日受付，9 月 7 日再受付）

松本　章代　（学生員）

2004 静岡大学大学院情報学研究科修士課程了．同大学院理工学研究科博士後期課程在学中．2005 東京工業高等専門学校助手，現在に至る．自然言語処理，情報検索に興味をもつ．情報処理学会会員．

小西　達裕　（正員）

1992 早稲田大学大学院理工学研究科博士後期課程了．1991 早稲田大学理工学部情報学科助手．1992 静岡大学工学部情報知識工学科助手．現在，同大学情報学部情報科学科准教授．知的教育システム，知的対話システム等に興味をもつ．博士（工学）．情報処理学会，人工知能学会，教育システム情報学会，日本認知科学会各会員．

高木　朗　（正員）

1974 早稲田大学大学院理工学研究科修士課程了．1974 同大学院博士後期課程編入．1981 早稲田大学大学院理工学研究科研究生．1983（株）CSK（現，(株) CSK システムズ）入社．2007 年 2 月退社．言語情報処理研究所設立，現在に至る．(独) 産業技術総合研究所客員研究員．自然言語処理等に関心をもつ．工博．情報処理学会，人工知能学会，日本認知科学会各会員．

小山　照夫　（正員）

1978 東京大学大学院工学系研究科産業機械工学専門課程了．工博．東京都老人総合研究所研究員，浜松医科大学助教授，学術情報センター助教授，同センター教授を経て，現在，国立情報学研究所教授．知識情報処理，データベース等の研究に従事．情報処理学会，人工知能学会等各会員．

三宅　芳雄

1974 東京大学大学院教育心理学専攻修士課程了．1982 Ph.D. in Psychology. カリフォルニア大学サンディエゴ校．国立教育研究所，日本電信電話（株）基礎研究所を経て，中京大学情報科学部認知科学科教授．現在，中京大学情報理工学部情報知能学科教授．認知科学の基礎，人の理解，学習過程の研究，ユーザビリティ研究に従事．情報処理学会，日本認知科学会，人工知能学会等各会員．

伊東　幸宏

1987 早稲田大学大学院理工学研究科博士後期課程了．同年早稲田大学理工学部電子通信学科助手．1990 静岡大学工学部情報知識工学科助教授．2000 静岡大学情報学部教授．現在，同大学創造科学技術大学院教授（情報学部兼務）．工博．自然言語処理，知的教育システム等に興味をもつ．情報処理学会，人工知能学会，言語処理学会，教育システム情報学会，日本認知科学会各会員．