静岡大学　　博士論文

検索キーワード間の
修飾－被修飾関係に基づく
WWW検索性能の向上

平成19年12月

大学院 理工学研究科
システム科学専攻

松本 章代

目 次

表 目 次

図 目 次

第1章　序論

本論文では，ウェブ検索エンジンの検索精度を更に向上させるため，ウェブ文書内における検索キーワード間の意味的係り受け関係を利用しようという試みについて述べる．

1 章ではまず，本研究の目的と背景，および本論文の構成について述べる．

1.1　研究の目的

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言いがたい．そこでウェブ検索エンジンの品質向上を目的として様々なアプローチが検討されているが，検索エンジンの基本的な戦略はユーザからの検索要求と適合するページを選択することである．ユーザの意図が最も反映され易いのはユーザからの入力であり，すなわち検索要求と文書との適合度が最も基本かつ重要であると考える．適合度の算出については，現状では語の出現に関する統計量を元に行われている．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．すなわち，統計的な手法で適合文書を求めることには限界がある．

適合度の算出には検索キーワードやページ内の意味に踏み込んだ処理を行うべきだと考える．そこで本研究では，複数のキーワードを用いて検索を行う場合において，それらのキーワードが文書内においてどのような構造で結びついているかに着目する．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

語が意味的関係をもって出現する文書を特定できれば，検索性能を向上させられることが見込める．

本論文では，検索キーワード間の意味的関係を表現しうる構造として「文構造」「表構造」「見出し構造」に着目し，ウェブ空間の多様な文書形式に対応する．各構造に応じた修飾－被修飾関係を利用して，検索対象文書における検索キーワードの意味的関係の強さ，すなわち適合度を推定する．

そこで，各構造ごとに，2語の検索キーワードが強い意味的な関係にあると推定される「係り受けパタン」を整理する．これに該当する場合は，適合ページであると判定する戦略をとる．

この提案手法の有効性を評価するために，既存の検索エンジンのフィルタリングツールとしてシステムを構築する.「係り受けパタン」に合致したらスコアを与え，降順に並び替える．上位の適合ページの量の変化を比較し，元にした検索エンジンの性能を向上させることに有効であることを実験的に検証する．

1.2 研究の背景

1.2.1 Web検索エンジンの現状

2007年現在，世界のWeb検索エンジンの勢力図は，Googleが50〜60%，Yahoo!が20〜30%，マイクロソフトのLive Searchが10〜20% となっており，この上位3つの検索エンジンによって占められている．

このような商用検索エンジンの目指す大きな方向は，広告収入である．すなわち，広告主から支払われた料金で検索結果ランキングを算出する．支払われた料金が多額であるほど高い順位を得ることができる．Overture（＝Yahoo!が買収）はこの料金が最優先される検索エンジンとして知られているが，Googleなどの他の商用検索エンジンも，他のアルゴリズムと組み合わせて利用している．そして多数の広告主を獲得するために利用ユーザを増やすことを目指す．そのためにはユーザの満足度を上げる必要がある．つまりユーザから評価されるサイトであるための取り組みとして，様々な検索結果の改善が行われている．

ランキングの問題点

企業のウェブサイトは，訪問者を増やすには検索エンジンで上位に表

示されることが不可欠だと考えている．そのために，多額の費用を使って，指定したキーワードで上位に表示してもらえる有料の検索サービスや，上位ランクの確保を指導するコンサルタントと契約する企業も多い．

一方，熱心なウェブログ運営者の多くは，とくに意図したわけでもないのに，自分のサイトが検索結果の上位に表示されるのを目にしている．検索のキーワードになった事柄について自分がとくに詳しいわけでもない，というケースも多い．

ウェブログ運営者たちは，検索結果で上位に表示される理由について，更新回数が多く，他のサイトからのリンクが多いからだと説明している．ほとんどの検索エンジンがランキングを決定する際に，この 2 つの要素を盛り込んでいる．

このことは，広告費の他，適合度（検索キーワードの出現頻度や係るタグの種類に応じた重み付け），引用度（リンク構造），新鮮度（ページ更新日時）などに基づく現状の検索エンジンのランキング手法が，まだまだ不十分であるということを端的に表している．情報提供者側の視点に立ってもランキングに対する意図が反映された結果であるとは言えないうえ，検索者側にとっても主観性の強いウェブログが上位にくることが満足の得られる検索結果だとは言い難い．

1.2.2 関連研究

現在，研究者間で行われているウェブ検索エンジンの品質向上研究のうち，ランキングスコアの改善に関する主な手法を以下に挙げる．

(1) 適合度

適合度とは「そのページの主題に検索語が適合しているページほど重要」という考えを基とした指標であり，この推計に用いられている主な検索モデルには

- ベクトル空間モデル（vector space model）
- 確率型モデル（probabilistic model）

などがある [1].

各モデルとも，最終的に文書得点を算出する必要があるため，語の出現に関する何らかの統計量を利用せざるを得ない．一般に，各モデルには TF と IDF という 2 種類の統計量が用いられている．TF とは term frequency の略で，ある文書に含まれているキーワードの頻度（出現回数）のこと，

IDF とは inverse document frequency の略で全文書のうちで，あるキーワードが出現している文書の割合，の逆数のことである．

ベクトル空間モデルでは，各文書及び検索要求における全単語（単語の総数を M とする）の重みを TF と IDF に基づいて算出し，M 次元ベクトルのなす角度を用いて類似度を定義している．確率型モデルでも TF と IDF を用いて各文書が検索要求と適合する確率を計算する．

しかし，TF-IDF 法には問題点がある．TF-IDF 法は，文書のテーマを表す重要な語は文書中に繰り返し出現するが他の多くの文書に出現する語は重要ではないという仮定に基づいている．各文書が新聞記事のようにある程度長い文書ならば，重要な単語が繰り返し出現するという TF 法の仮定は有効である．しかし，検索対象の文書が十分な長さがない場合は，例え重要な語でも，他の語と比べ使用頻度に明らかな差は現れにくい．すると，重要な語とそうでない語の TF に差がでない．このようなケースでは，IDF だけで単語重要度を決めることになる．よって TF-IDF 法では高精度な文書検索は期待できない．

そこで，一般的には，キーワードの出現位置や近接度，タグの種類による重み付けなどによって補正が行われている．

(2) 新鮮度

ページの更新日時の新しいものほど，高いスコアを付ける．同じような内容を取り上げていても，新しいページの方が，新たに得られた情報や議論の結果を取り込んで情報の価値や量が高まっている可能性が高い．

(3) 引用度

Google の PageRank に代表される引用度スコアは，多数引用されるページは信頼できる，また，信頼できるページに引用されるページも信頼できる，という考えに基づいたものである．

WWW においては，ハイパーリンク構造を解析することによって求められる．この手法は次の 4 種類に大きく分類できる．

(3-1) Link Popularity

Link Popularity は，数多くリンクされている Web ページほど重要だとみなす手法である．例えば，検索された Web ページの被リンク数や，検索された Web ページが存在するサーバ全体の総被リンク数の多い検索結果のスコアを高くする手法が使われている．

(3-2) HITS

Kleinberg らが提案した HITS(Hyper Induced Topic Search）は，ある特定のトピックに関する情報源であるオーソリティ (authority) と，オーソリティへのハイパーリンクの集合であるハブ (hub) という 2 種類の Web ページに対して，良いオーソリティは多くの良いハブからリンクされ，良いハブは多くの重要なオーソリティをリンクすると言う相互依存する関係を求めることで，検索結果の質を改善する手法である [2].

(3-3) PageRank

Page らが提案した PageRank は，多くの良質な Web ページからリンクされている Web ページは，良質な情報源であると考え，Web のリンクをランダムにたどる”random surfer”行動モデルに基づいて，Web ページが閲覧される確率を検索して得られる Web ページの重要度である [3, 4].

PageRank の基本的な仕組みは，あるページの PageRank を，そのページに存在する発リンク数で割った数が，それぞれ被リンク先の PageRank に加算されるということである．すなわち，PageRank を高くするためのポイントは，大きく分けて 3 つある．

- 被リンク数 (単純な意味での人気度の指標)
- お勧め度の高いページからのリンクかどうか (裏付けのある人気かどうかの指標)
- リンク元ページでのリンク数 (選び抜かれた人気かどうかの指標)

まず基本的に，多くのページからリンクが張られていればお勧め度は高くなる.「(たくさんリンクされるような) 人気のあるページは，きっと良いページであるに違いない」ということである．被リンク数を人気度の指標の一つと見ることは自然な考え方であろう．リンクを張るという行為は,「このページを見るといい／このページは役に立つ」という推薦行為を行っていることとみなせるからである．だが，PageRank の考え方はそれだけにはとどまっていない.

すなわち，単に被リンク数の多寡だけではなくお勧め度の高いページからのリンクは高く評価する．また同時に，総リンク数が少ないページからのリンクは高く評価し，総リンク数が多いページからのリンクは低く評価する．言い換えれば「(多くの推薦を集めるような) 良いページが推薦するページは，同じく良いページであるに違いない」という判断と,「リンクを過剰に乱発するインフレ気味なリンクに比べて，選び抜かれたリンクは良質なリンクであるに違いない」という判断を同時に行っている.

従来は，ページの重要度としてそのページの被リンク数だけを単純に用いることがあったが，PageRank 方式の場合，機械的に生成されたリンクの影響を受けにくいという利点がある．つまり，PageRank を上げるためには良質なページからリンクされる必要がある．

(3-4) Cocitation アルゴリズム

Dean らが提案した Cocitation アルゴリズムは，何らかの関連を持つ Web ページをリンク解析で発見する手法として，二つの Web ページを同時に引用している Web ページ群を求めて，その数により関連性が高い Web ページを発見する手法である [5].

(4) 人気度

多数の利用者が参照したページほど重要という考えに基づき，スコアを与える．人気度は過去の参照履歴から算出できる．すなわち，ある検索キーワードと，その検索結果から利用者が選択したページ（ジャンプ先 URL）の組を記録しておけば，各検索キーワードに対してどのページへのジャンプ回数が多かったかがわかる．この過去のジャンプ回数が，検索キーワードごとの人気度に相当する．

大久保らは，検索エンジンで検索された語の頻度から，利用者全体の情報需要のトレンドを分析する方法を示した [6]．この方法では，同一利用者が短い時間内に使用した検索語は同じ目的で利用されているという仮定に基づいて，例えば「桜」と「花見」のような単語をグループ化して扱うことで，類義語が別々に集計されないように工夫している．

1.3 本論文の構成

本論文は全 7 章で構成される．

まず 2 章で，本研究の着眼点及び 2 語の修飾－被修飾関係を用いることによるカバレッジについて説明する．次に 3 章で，各構造ごとに検索キーワードの修飾－被修飾関係を抽出する手法を議論する．次の 4 章では，3 章で議論した方法を具現化するための詳細な手段を紹介する．実装した試作システムを用いて，5 章では，その性能を評価し，類似手法と比べて効果があることを確認したのでそれを述べる．さらに 6 章で，精度向上に向けた現在の取り組みについて報告する．最後に 7 章で，成果と今後の課題について述べ，締めくくる．

第2章　基礎的考察

2章では，本研究の着眼点について述べ，2語キーワード間の修飾－被修飾関係とは何かを説明し，さらにその適用範囲を検証する．

2.1　本研究の着眼点

1.2.2節で述べたように，主なものだけを取り上げてみてもWeb検索エンジンの品質向上には様々なアプローチが考えられるが，検索エンジンの基本的な戦略はユーザからの検索要求と適合するページを選択することである．ユーザの意図が最も反映され易いのはユーザからの入力であり，すなわち検索要求と文書との適合度が最も基本かつ重要であると考える．

しかしながら，適合度の算出については先述の通り，現状では統計量を元に行われているが，統計的な手法で適合文書を求めることには限界がある．

本研究では，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ついているかに着目して，検索精度を向上させることを試みる．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ている．従って，それらの語が文書中に存在するということだけでなく，それらの語の間の修飾－被修飾関係を表現しうる構造を形成している文書を特定することにより，検索精度の向上が期待できると考えている．

修飾－被修飾関係という係り受け構造を手掛かりとして文書検索を行う研究としては，文献 [7, 8, 9, 10, 11, 12, 13] など数多く存在する．

TREC-5では，Strzalkowskiら [7] により，ヘッドの語と修飾関係にある語のペア（V+OやS+V）でインデクシングを行なうことにより，精度の向上が見込めることが報告されている．

新美ら [8, 9] は，特許データを対象として，単語間の係り受け関係を利用した全文検索システムを構築し，ブール型検索や近接関係を用いた検索より有効であることを示した．クエリを自然言語ではなくキーワード対で入力させる等，本研究と類似している面もある．ただし新美らの手法では，指定したキーワード対が，検索対象の文中に直接係って出現する必要があるため，再現率を落とす可能性が高い．

峯ら [10, 11] は，クエリを自然言語で受け付け，構文解析を行い，係る単語と係られる単語のペアの集合を作成して検索対象のテキスト集合から単語のペアを含む文を抽出し，係り受け関係の一致の度合いによって適合度を判定する手法を用いた．清田ら [12, 13] は，自然言語を受け付ける質問応答システムを作成した．これらはいずれも，クエリを自然言語文で受け付け，それと同じ係り受け構造を含む文書を抽出するという手法である．そのため，クエリと同等の意味を異なる係り受けで表現している文書は検索できず，再現率を大きく落とす危険性を否定できない．この弱点を補うため，峯らは連体助詞句を伴う名詞句（「芥川龍之介の本」）と連体助詞句を省略した表現（「芥川龍之介本」），一方の名詞が動詞を伴って連体修飾節を構成して他方を修飾するパタン（「芥川龍之介が書いた本」）とを同等と見なせるようにルールを設定している．また，清田らでも同様に，名詞Ａが格助詞を介して動詞に係るケースと，名詞Ａが連体名詞を介して名詞Ｂに係り名詞Ｂは格助詞を介して動詞に係るケースを同等とみなす，といった言い換えに対応する仕組みを提案している．

これに対し本論文では，2 語のキーワードから推定される検索意図を表す係り受けとして妥当なものを推定して検索を行うという方法をとる．その際，2 語を結びつけうる係り受け構造について考察し，可能な係り受けパタンをできる限り網羅的に整理することにより，言い換えに対する頑健性の向上を図る．更に，文中における位置を用いた判定を加えることにより，精度の向上を目指すが，こうした試みは他の研究では未だ検討されていない．

本論文では，検索に際して複数のキーワードが入力される状況を想定する．以降，クエリとはこの検索キーワードの集合を指す．また，検索キーワードは，2 語に限定して考えることとする．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み合わせとして捉えることができること，Jansen ら [14] や風間ら [15] 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．

また，ウェブ空間の文書形式は多様であるため，キーワード間の修飾－被修飾関係が文の形で表現されるとは限らない．表構造および見出し構造もまた，HTML 文書中でキーワード間の修飾－被修飾関係を表しうる．そこで本研究では，検索キーワード間の修飾－被修飾関係を抽出するために，文・表・見出しの各構造に応じた修飾－被修飾関係を利用して，検索対象文書における検索キーワードの意味的な結びつきの強さ，すなわち適合度を推定する．

そこで，各構造ごとに，2 語の検索キーワードが強い意味的な関係にあると推定される「係り受けパタン」を整理する．これに該当する場合は，適合ページであると判定する戦略をとる．以下に，各構造ごとの「係り受けパタン」を説明する．

(1) 文構造

2 つのキーワードが両方とも 1 つの文に含まれていた場合は，そのキーワード間の修飾―被修飾関係を利用して，検索意図に適ったページであるか否かの推定が可能である．

例えば,「人気　ノート PC」というキーワードで検索を行うとき，ページ中に「2004 年度の人気ノート PC」という文が含まれる場合と「人気ノート PC ケースを販売中」という文が含まれる場合とでは，前者が検索意図に近いと推定できる．前者は「人気」が「ノート PC」を修飾しているのに対し，後者は「人気」が「ケース」を修飾している（図 2.1）．すなわち,「人気」があるのは「ケース」であって「ノート PC」ではない．このように，1 つの文に存在する 2 つの検索キーワードが強い関係で結ばれていることが，検索意図に適ったページを見つける手がかりとなりうる．

ただし，2 語が直接あるいは他の語等を介して修飾－被修飾関係をもつ際の係り受け構造は，2 語の品詞や語の意味の組み合わせによって異なる．そこで，2 語の品詞や意味分類の組み合わせごとにその間に想定可能な係り受け構造を整理する．検索キーワードの意味分類に応じて，その組合せによって抽出すべき係り受け構造を制限し,「係り受けパタン」を決める．

(2) 表構造

表構造は，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．表の構成要素の中の 2 つ

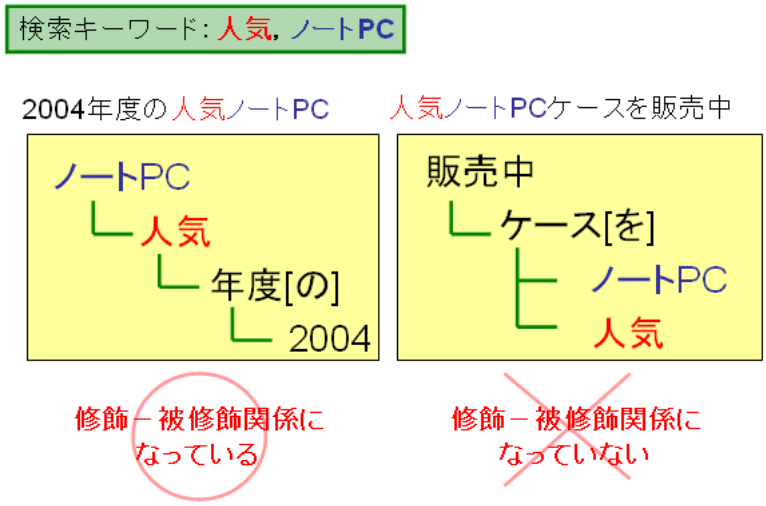

図 2.1: 文構造

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 2 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）修飾－被修飾関係と，表の構成要素間に想定される修飾－被修飾関係とが一致する場合，修飾－被修飾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．

例えば,「マレーシア　首都」というキーワードで検索を行うとき，列の見出しに「マレーシア」，行の見出しに「首都」と書かれている表があれば，その見出しが直交するセルに答えが書かれている可能性が高い．一方，1 つの表に「マレーシア」と「首都」が存在していたとしても，それらが互いに無関係な位置に出現する場合は，検索者が求める情報が表の中に存在する可能性は低い（図 2.2）．

すなわち，表を構成する修飾－被修飾関係の性質を，適合／不適合の判定に利用することができる．表の構成要素間に修飾－被修飾関係を認めることができる組み合わせから推定し「係り受けパタン」を決める．

(3) 見出し構造

例えば,「チーズケーキ　作り方」というキーワードで検索を行うとき，大見出しに「チーズケーキ」，その中の小見出しに「作り方」と書かれて

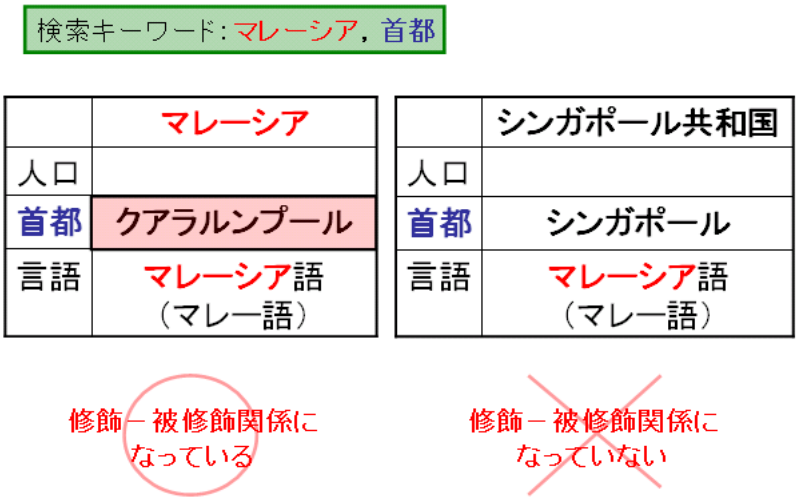

図 2.2: 表構造

いるページには，チーズケーキの作り方が書かれているであろうことが容易に予測できる．一方，小見出しに「作り方」と書かれていても，「チーズケーキ」が「作り方」の見出しになっておらず，あるいは小見出しが影響を及ぼす範囲内に「チーズケーキ」が存在しなければ，2 つのキーワードの関連は薄い（図 2.3）．

そこで，ページ全体をツリー構造（階層構造）としてとらえる．一方の検索キーワードが見出しに含まれているとき，その階層下にある見出しや箇条書きの中にもう一方の語が含まれていれば，修飾－被修飾関係にあると見なし，これを「係り受けパタン」とする．

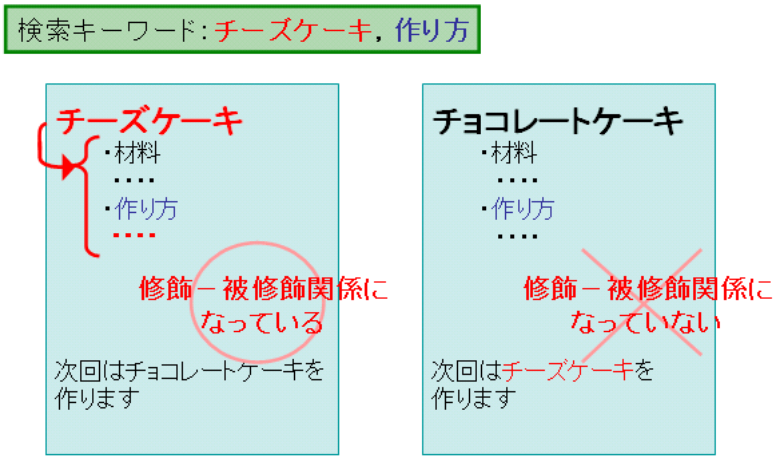
検索キーワード：チーズケーキ，作り方
チーズケーキ
・材料
・・・・
・作り方
・・・・
修飾－被修飾関係に
なっている
次回はチョコレートケーキを
作ります
チョコレートケーキ
・材料
・・・・
・作り方
・・・・
修飾－被修飾関係に
なっていない
次回はチーズケーキを
作ります

図 2.3: 見出し構造

2.2 実験データに基づくカバレッジ分析

実際のウェブ検索において，2 語の修飾－被修飾関係を用いることによってどのくらいの範囲をカバーできるかを検討する．まず，ウェブ検索エンジンに入力されるキーワードの傾向から本研究で対象とする範囲を確認する．また，既存の検索エンジンで拾い上げてしまう不適合ページに見受けられる特徴を検討して，本論文で提案する手法が精度の向上に有効な範囲について考察する．

2.2.1 クエリの分析

情報学部の大学生約 20 名に依頼し，一定期間，日常生活の中で実際に検索サイトに入力したデータを収集する実験を行った．計 1126 件のデータ中，1 語で検索が行われたケースが 42.9%，2 語は 39.6%，3 語以上は 17.5%という構成であり，2 語で検索されたケースからランダムに抽出した 200 件について 2 語の関係性を調査した．その結果，2 語が修飾－被修飾の関係にある[1]ケース 86%，同義語・類義語の関係にあるケース 5%，その他 9%であった．なお，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）はその他のケースに分類している．

本論文では，2 語のキーワード間の修飾－被修飾の関係を利用することから，2 語以上のキーワードを必要とする程度に複雑な検索の 60%，全体の 34%が本論文の適応可能な範囲となる．

2.2.2 不適合要因の分析

次に，一般的な検索エンジンにおいて，検索精度（precision）を下げる要因について，以下の手順で分析を行った．

[1]一方が他方を修飾する関係において自然な句・節を作れる場合に修飾－被修飾関係と判断した．なお，2 語が修飾－被修飾関係にあると判断できる場合でも，検索者が直接的な修飾－被修飾関係を意識せず，全く別の意図で検索する場合も，原理的にはありうる．しかし，紙数の都合上詳細は割愛するが，我々の予備実験によると，そのようなケースは 2%弱（170 件中 3 件）とまれである．

(1) 「USJ に行きたい」「ノートパソコンを購入したい」などの状況を 4 つ決め，それらの状況に対し「USJ のチケットをあらかじめ浜松で購入しておきたい」「(USJ の) 近くの安いホテルに泊まりたい」といった，さらに具体的な検索課題を各 10 個，計 40 個設定し，その際どのようなクエリを入力するかについて，アンケート調査を実施する．

(2) 2 語のクエリで修飾－被修飾関係にあるものの中から課題の重複が無いよう 10 個をランダムに抽出し，その 2 語で実際に検索を行う．

(3) 各々上位 100 位までについて適合／不適合の判定を人手で行う．

(4) (3) で収集された全ての不適合ページのうち上位の方から約 200 ページを抽出し，不適合要因を分析する．

その結果，不適合ページは以下のように大別できることが判明した．なお，ここで「記述単位」とは，1 つの文・1 つの表・1 組の見出しを指す．

(1) 2 つのキーワードが同一の記述単位に含まれていない，あるいは並列な関係にある．

(2) 2 つのキーワードが同一の記述単位に含まれている．

(2-1) 2 つのキーワードが修飾－被修飾関係にある．

(2-1-1) キーワードが文書全体の話題の中心とはなっていない．

(2-1-2) キーワードが多義語であり想定外の意味で使われている．

(2-1-3) 他の語の影響を受け，検索意図から外れてしまっている．

(2-2) 2 つのキーワードが修飾－被修飾関係にない．

今回対象とした約 200 ページについては (1) および (2-2) が全体の約 30%，(2-1-1) が約 25%，(2-1-2) が約 5%，(2-1-3) が約 40%を占めるという結果になった．

(2-1-1) については，主に 2 つの傾向が見受けられた．一つは，(a) 文や名詞句の主題に検索キーワードが含まれていないパタンである．例えば「世界　電圧」というキーワードのとき「世界の電圧に対応したトラベルクッカー」の紹介ページは不適合である．この場合は，文の中の係り受けの位置を考慮することによって対処が可能である（3.1.3 節参照）．もう一つは，(b) 文書全体に対して検索キーワードに関して記述されている

部分の扱いが小さいパタンである．文書中に「世界の電圧はまちまちだ」としか書かれておらず，具体的な情報が得られないケースなどである．

(2-1-3) の「他の語の影響を受け，検索意図から外れてしまっている」とは，他の語を修飾したり，他の語によって修飾されることによって，検索意図から外れてしまったケースである．例えば「自動車　イベント」という検索キーワードのとき「自動車学校のイベント」のページは検索者の想定外であろう．

本研究では，キーワード間に修飾－被修飾関係が有るものを判定するための仕組みを作り，(1) および (2-2) のパタンを排除する．さらに，(2-1-1) についても対応を極力検討する．これよって 5 割以上の不適合ページが排除できるものと期待できる．

しかしながら，検索エンジンとしての有効性を確認するためには再現率に関する議論も不可欠であり，精度の議論だけで一概に結論を導くことはできない．本論文では，高い再現率を維持しつつ精度を下げる要因を極力排除することを目指す．

第3章　修飾－被修飾関係を用いたページフィルタリング手法

3章では，修飾－被修飾関係を用いたページフィルタリングの手法について，文・表・見出しの各構造ごとに議論する．

3.1　文構造による係り受け関係

3.1.1　修飾－被修飾関係を用いることの妥当性

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，同義語・類義語の関係を考えることができる．同義関係にある2つの語は，一方の語が同格的に他方の語に係り，2語が結び付けられる構造を構成する．この2語がキーワードに指定されるのは，1語でも検索したい対象を比較的十分に特定しうるが，それを表す語彙が複数想定される場合に多用される．例えば,「サーチエンジン」と「検索エンジン」の2語をキーワードに指定する場合などがこれにあたる．

それ以外の意味的関係をもつ2語の場合，文構造の係り受け関係は，以下の構造のいずれかに分けられる（図3.1）．

(1) 自立語を介さず直接あるいは助詞等の付属語のみを介して一方が他方を修飾する．

(2) 自立語を介して係る．

(3) 別の自立語に双方の名詞が直接あるいは間接的に係る．

(3) は，途中に接続助詞や引用の「と」，連体節を構成する「という」などを含む場合などいくつかの例外を除き，一方の語がヘッド[1]になるように係り受け構造を組み替えて (2) の構造に再構成することが可能である．例えば「パソコンをネットショップで購入できる」における「パソコン」と「ネットショップ」は (3) の構造であるが,「パソコンを購入できるネットショップ」と (2) の形で言い換えることができる（図 3.2）．例外的な場合，すなわち，接続助詞，引用の「と」，連体節を構成する「という」などを介した場合，それに係る句・節と，係られる句・節の関係は語と語の係りではなく，句・節との関係付けとなる．従って，それらを介して 2 つのキーワードが結び付けられている場合，両者は直接係り受け関係で結ばれるわけではない．そこで，これら例外的な場合は無視し，(2) に変換可能な (3) の構造，および (1)，(2) の構造に該当する同義語・類義語の関係に無い 2 語を，広義の「修飾－被修飾関係」と呼ぶ．(1)，(2) の構造をとる 2 語がキーワードとして用いられるのは，1 語だけでは意味が一般的すぎて検索したい対象を十分に特定できず，更に 1 語を追加して検索対象を限定する必要がある場合である．このとき，2 語のキーワードは，上述の (2),(3) のように他の自立語を含む場合であっても，その 2 語を含み，かつユーザが希望する検索対象を十分限定できるような句・節・文を構成できるはずである．その句・節・文によって妥当な程度に具体的，限定的に検索対象を表現できるような語の組として，2 語のキーワードは選択されると考えられる．

このような修飾－被修飾関係にある 2 語が現実の検索時においてキーワードとして用いられることが多いと考えられるため，本論文では，修飾－被修飾関係に着目して検討する．キーワードが修飾－被修飾関係にある 2 語として選択されているのであれば，その 2 語から再現できる句・節・文構造あるいはそれと等価な構造を含む文が存在する文書中に，検索対象が具体的に記述されている可能性が高い．よって，2 語が修飾－被修飾関係となって出現する文が存在することを，該当ページが適合であると判定をするための第 1 の条件として設定する．

さらに，修飾－被修飾関係にある 2 語が構成する句や節が，それらを含む文や名詞句の主題[2]を構成する要素である場合，その文自体が検索対象に関する記述であったり，そうでなくとも前後の文脈の中に検索対象

[1]ヘッドとは文（または節や句）を依存構造木にしたときルートとなる語のことであり，日本語の場合は文（節・句）末の自立語となる．

[2]本論文では，主節や体言止めの名詞句のヘッドといった文や句の中心要素を「主題」と呼ぶ．

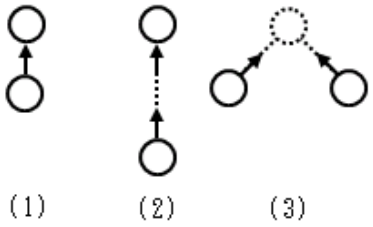

図 3.1: 2語の構造

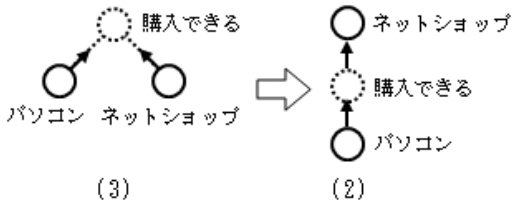

図 3.2: 2語の再構成

に関する記述があったりすることが期待でき，検索精度を向上させることが可能と考えられる．修飾－被修飾関係にある2語が構成する句や節が文や名詞句の主題を構成するか否かは，その2語の文中における位置を手掛かりに判定可能である．そこで，2語が文や名詞句の主題を構成する位置に存在することを,（第1の条件を満たす文書をさらに限定するための）第2の条件とする．

すなわち，本論文では，2語キーワードが修飾－被修飾の関係にあるとみなし，その2語が修飾－被修飾の関係をもち，さらにその2語が文や名詞句の主題を構成する位置にあるような文が出現する文書を優先させるという方針で，ウェブ検索の精度向上を目指す．

3.1.2　2語キーワードが構成する係り受け構造に関する考察

2語が直接あるいは他の語等を介して修飾－被修飾関係をもつ際の係り受け構造は，2語の品詞の組み合わせによって異なる．そこで，2語の品詞の組み合わせごとにその間に想定可能な係り受け構造を整理する必要がある．しかし，2語で検索する際に用いられる語のほとんどは名詞であ

る．そこでまず，名詞 2 語の場合について検討を行い，その後で他の品詞の語がキーワードとして用いられる場合について検討する．

名詞 2 語のキーワードの場合

本項では，名詞の意味分類と検討対象とする 2 語の組み合わせについて検討し，次いで個々の組み合わせごとに 2 語の間に想定すべき係り受け構造について議論する．

語と語の間の係り受け構造のあり方は，基本的には品詞によって定まるが，更に，2 語の意味の組み合わせによって検索対象を表現するための可能な係り受け構造が制約される．例えば,「中華料理」「野菜」という 2 語の場合,「野菜の中華料理」「中華料理の野菜」のようにどちらが被修飾語となる係り受け構造も考えることができ，それぞれ「野菜を主材料とする中華料理を調べたい」「中華料理でよく使われる野菜を調べたい」という検索意図を推定することができるが,「車」と「100 万円」という 2 語の場合,「100 万円の車」の検索意図は容易に推定できても,「車の 100 万円」では何を検索したいか想像しがたい．

そこでまず，名詞を意味カテゴリに分類して検討する．この分類は，世界が「もの（実体）」と「こと（現象）」から構成されると捉え，それぞれが「属性」をもち，個別の「属性値」を取ることにより意味が特定されると考えられることに基づく分類である．

- 実体を表す名詞（車，Linux，宗教など，以下「実体名詞」）
- 現象を表す名詞（インストール，検索など．いわゆるサ変名詞はここに分類される．以下「現象名詞」）
- 属性の名称を表す名詞（料金，色，方法など，以下「属性名詞」）
- 属性値を表す名詞（3776m，赤など，以下「値名詞」）

意味カテゴリをこの 4 つに分けることで，その組み合わせによって生じる係り受け構造のバリエーションが限定される．この名詞の分類は，Takagi ら [16] によるものに基づき，それを改変したものである．一方，従来から言語学の分野では，様々な名詞の分類手法が提案されている [17]．例えば，名詞全体を，普通名詞・固有名詞・集合名詞・物質名詞・抽象名詞の 5 つに分類する手法 [18] を，我々の分類と大筋で対応付けると以下のようになる．

① 普通名詞：一定の形や大きさをもつ物体．··· 実体名詞

② 固有名詞：人・場所・製品などの名前．··· 実体名詞

③ 集合名詞：同じ種類の人や物の集合体．··· 実体名詞

④ 物質名詞：一定の形や大きさのない物質．··· 実体名詞

⑤ 抽象名詞：形がなく，目に見えない性質・動作など．··· 現象名詞，属性名詞，値名詞

これらの従来の分類と比べ，本研究の分類に基づくことで，係り受け構造の制約の利用という観点から，存在可能な名詞の組み合わせとそうでない組み合わせの弁別をより効果的に行えると考える．例えば，従来の分類では,「1kg」「重さ」はいずれも抽象名詞にあたるため「カメラ」と「1kg」という組み合わせと「カメラ」と「重さ」という組み合わせとを区別することができない．一方本研究の分類に基づけば,「1kg」は値名詞,「重さ」は属性名詞と分類される．実体名詞と値名詞の組み合わせの場合は「値名詞の実体名詞（1kg のカメラ）」，実体名詞と属性名詞の組み合わせの場合は「実体名詞の属性名詞（カメラの重さ）」という係り受け構造が，検索対象の表現として存在しうるのに対し,「実体名詞の値名詞（カメラの 1kg）」「属性名詞の実体名詞（重さのカメラ）」は検索対象の表現としては存在しないものとして排除できる．また，より詳しい名詞分類として EDR 電子化辞書[3]で採用されている分類がある．そこでの分類と本論文で用いる分類とを対比させると，おおよそ表 3.1 のような対応関係になる．この対応表から明らかなように，EDR 概念辞書の最上位レベルの 5 分類と本論文で採用している 4 分類の間に単純な対応関係はない．例えば，EDR の「ものごと」の中には本研究における現象，属性，実体に対応するものが含まれており，本研究で採用した分類に基づく係り受け構造の制約を表現することはできない．我々の研究においては，係りの性質が規定できる，安定した分類であることを重要視している．現象・実体・属性・値という 4 つの概念の分類は，安定した意味分類であると共に，文の最も基本的な依存構造のあり方と密接に関係する分類法であると考えている．より詳細な意味分類を用いると，語の意味は状況や立場などによって様々に変化することから，状況や立場に応じた意味分類に注力する必要が生じかねない．そこで，本論文では，この 4 分類のレベ

[3] http://www2.nict.go.jp/r/r312/EDR/J_index.html

表 3.1: 概念辞書の対応

EDR	本手法
1 人間または人間と似た振る舞いをする主体	実体
2 ものごと	
2-1 もの	実体
2-2 事柄	現象
2-3 識別名	属性
2-4 客観的な対象	実体
3 事象	
3-1 現象	現象
3-2 行為	現象
3-3 移動	現象
3-4 変化	現象
3-5 状態	
属性名	属性
値	値
その他	値
4 位置	実体
5 時	値

ルで語の組み合わせによって抽出すべき係り受け構造を制限するという研究方向の有効性を検証する．

これら 4 種類の名詞の組み合わせは単純に考えれば 10 通りあるが，我々が収集したウェブ検索エンジンのログ（2.2 節参照）において 2 語で検索が行われたケース（500 件程度）を調査したところ，実体名詞を少なくとも 1 つは伴うケースが全体の 96.1%を占めた．このような結果となった理由を以下で考察する．

まず，検索キーワードとして現象名詞が用いられる場合を考える．一般に，現象名詞 1 語では個別の現象を特定することができないことが多い．例えば「インストール」という 1 語では，漠然としすぎて何を検索したいかを伝えることはできない．このため，検索条件となりうる程度に具体的な現象を特定しようとした場合，現象にかかわる実体を併せて指定する必要がある．従って，現象名詞は実体名詞を伴って検索条件と

なる可能性が高い.

属性名詞は，その属性を内包する実体に言及せずに，単独で検索条件となることは考えにくい．実体名詞と共に指定された場合，その実体の該当する属性値を知りたいという意図を想定することができる．属性名詞を値名詞と共に2語で用いられる場合は，指定された属性値をもつ実体を検索したい場合と考えられるが，属性値だけでは検索したい実体が多岐にわたりすぎ，現実的な検索条件としてはあまり適切ではない（高さが3776mである山を知りたい場合,「高さ　3776m」よりも「3776m　山」とする方が自然である）.

値名詞も単独では検索条件とはなりにくく，上の例のように（固有名詞ではなく，クラスを表す）実体名詞と共に用いられて，指定された属性値をもつ実体を検索するときに用いられる.

そこで，名詞2語の組み合わせを

- 実体名詞＋現象名詞（例：新幹線　予約）
- 実体名詞＋属性名詞（例：カメ　寿命）
- 実体名詞＋値名詞（例：100円　ラーメン）
- 実体名詞＋実体名詞（例：USJ　チケット）

の4組に限定して，想定すべき係り受けパタンを検討し，整理することにする[4]．結果は後出の表3.2にまとめる．以下の検討における(a)〜(f)は，各々表3.2中の記号に対応する.

A) 実体名詞＋現象名詞　ここでは現象を表す語も名詞として用いられる場合について可能な係り受け構造を検討する．ただし，サ変名詞が「する」を伴って動詞として用いられる文を含む文書中も拾い上げるため，この組み合わせの場合には，以下で述べる2語の名詞の場合の処理に加え，現象名詞に「する」を補い，後述する「実体名詞＋動詞」の処理も行う.

この組み合わせにおいて，実体名詞は現象名詞を用言化した場合の格名詞[5]となる．ゆえに，実体名詞が現象名詞を修飾する場合は，現象名詞

[4]現象を表す語には「エルニーニョ」「天安門事件」など具体的な特定の現象を表すものもあり，この場合は実体を伴わず，例えば「エルニーニョ・原因」のように2語キーワードを構成することもある．将来的にはこのようなケースにも対応できるよう,「現象名詞」＋「属性名詞」の組み合わせも検討する必要がある.

[5]格助詞を介して述語に係る名詞を「格名詞」と呼ぶ.

に，格助詞が転化したタイプの連体助詞「の」「からの」「への」「での」などを介して実体名詞が係るという依存構造を構成する (b)．また，連体助詞を省略し実体名詞が直接現象名詞を修飾する場合もある (a)．逆に現象名詞が実体名詞を修飾する場合は，現象名詞が実体名詞を修飾する連体修飾節中で用いられる場合（「加湿 ができる エアコン」など）(c)，現象名詞が直接実体名詞に係る[6]場合（「加湿 エアコン」など）(a)(b) がある．

B) 実体名詞＋属性名詞　この組み合わせにおいて，属性が実体属性の場合は，実体が属性を内包するという関係をもつ．例えば「スイカ」と「糖度」ならば,「スイカが糖度を内包する」という関係をもつ．また属性が現象属性の場合は，実体と属性はその現象概念を介して関係付けられる．例えば,「新幹線」と「速度」は「走る」という現象を介して「新幹線が,〜の速度で走る」のように関係付けられる．このような関係を想定した場合，検索キーワードとして実体名詞と属性名詞の 2 語が指定された際の検索したい内容を表す表現として，属性名詞をヘッドとし実体名詞が連体修飾を構成する構造（例えば「スイカがもつ糖度」「新幹線が走る速度」など）を考えることができる（実体名詞をヘッドとする構造，例えば「（特定あるいは不定の）糖度をもつスイカ」が検索したい対象を表しているケースは考えにくい）．そのような検索意図を仮定すると,「高い糖度のスイカの栽培法」のように実体名詞をヘッドとする文が見つかっても，必ずしも必要な情報を提供しているとは言い難い．そこで，この組み合わせの場合，実体名詞をヘッドとする係りは対象外とする．

そこで，実体名詞が現象を表す述語を用いて連体修飾節を構成し属性名詞を修飾する構造を基本形 (c) として捉え，これをベースに言い換えのバリエーションを検討する．属性名詞が実体属性・現象属性のいずれであっても，基本形は属性名詞をヘッドとする連体修飾節構造であり，さらに連体修飾節から実体名詞を除いた部分（上述の例で言えば「がもっている」「が走る」の部分に相当する.）の意味を連体助詞「の」などで言い換えることができる．従って「スイカがもっている糖度」は「スイカの糖度」,「新幹線が走る速度」は「新幹線の速度」と言い換えられる (b)．また連体助詞を省略する構造もありうる (a)．また，内包の述語を主動詞とし，実体名詞と属性名詞とが格名詞となって文を構成し，その文や前後の文脈の中で必要な情報を提示している可能性がある（例えば,「スイカは 10〜13 程度の糖度をもつ.」など.）(d)．

[6]厳密には，現象名詞が実体名詞に直接“ 係る ”ことはない．連体修飾節構造を介している（省略している）．

C) 実体名詞＋値名詞　この組み合わせの場合，ある属性の値が指定された属性値と等しい実体が検索対象であると考えられるため，実体名詞をヘッドとして，値名詞が連体修飾を構成する係りを考えればよく，逆を考慮する必要はない．実体名詞と値名詞の場合，B) で述べた関係で実体と属性が関係付けられ，更に属性と値とが「属性は値に等しい」という関係で結ばれる形が基本形となる．従って,「値」に等しい属性を内包している「実体」（例えば「3000m に等しい高さを内包している山」）という連体修飾節構造がベースとなり，そこから実体名詞と値名詞を除いた部分（上述の例で言えば「に等しい高さを内包している」）の意味を連体助詞「の」などで言い換えることができる（さらにその連体助詞を省略することも）(a)(b).

ここで，値名詞と実体名詞とは「等しい」と「内包する」という2つの述語及び属性名詞を介して接続している．従って，値名詞を直接格名詞に取る述語が連体修飾節を構成して実体名詞を修飾する係りは想定にしくい[7]．また，一つの述語に実体名詞と値名詞とがともに格名詞として接続する構造も基本的にはない．しかし，実体名詞が提題化され，提題助詞「は」を介して述語「等しい（である）」に係ること（「富士山は高さが 3776m である.」など）は考えられるため，このパタンも追加しておく (d).

D) 実体名詞＋実体名詞　2語ともに実体の場合は，2実体を結びつける現象を表す用言に2つの名詞が格名詞として接続する構造（「万年筆はインクを必要とする.」など）が想定できる (d)．ここで，一方の実体名詞は，検索対象を表す他方の実体名詞の意味を，より限定するために用いられているはずである．従って一方の実体名詞は用言と共に連体修飾節を構成する（「万年筆が必要とするインク」など）(c)．その場合，2つの実体を挙げただけでその関係（媒介となる現象）が自明である場合には，用言を省略して連体助詞で結んだり（「万年筆のインク」など），間に語を介さず直接係る（「万年筆インク」など）等の文構造が取られる可能性がある (a)(b)．また両者が対等であるので相互がヘッドとなるパタンをそれぞれ考える必要がある．

---

[7]「高さが 3000m である山」のように，英語の所有格関係節に相当する文構造で連体修飾節を構成することは考えられるが，現在のこのパタンは対象としていない．また「車が 200 台入る駐車場」のように，数量格成分として値が係るものもあるが，現在のところ対象から除外している．

E) 対象とする係り受けパタン 2語の意味の組み合わせから決まる係りの基本形とその言い換えのバリエーションが本手法の係り受けパタンの基本であるが，それに当てはまらない例外的なパタンも存在する．それらは予備実験の際に数種類見受けられ，現在のところはその中で頻出したパタンについてルール化を行った．それを (I)(II) として以下に示す．

(I) 現象名詞がある名詞に連接し句を構成している場合,（現象名詞とそれが係る名詞が強く結びつき，1語の複合語のように解釈することができるため,）それに対して意味的には現象名詞に係るべき実体名詞が連体助詞を介して修飾しようとするとヘッド側の名詞に係ると解析される．例えば,「Linux のインストール方法」において「Linux の」と「インストール」とがともに「方法」に係り,「Linux」は「インストール」に係らない．しかしこれは「Linux をインストールする方法」を簡便な名詞句構造に同義変形したものと捉えることができる．従ってこの場合，実体名詞は現象名詞に係るとみなすことができる．

(II) 対象文書内で，表層上2つのキーワード（名詞1，名詞2とする）の間に別の名詞（名詞3とする）が挟まり，それが一方のキーワードとともに複合語を構成する場合がある．例えば「マレーシア　大使館」「LZH　解凍」といった検索キーワードに対し，実際のウェブページ中には「マレーシア日本大使館」「LZH ファイルの解凍」のように出現するケースがこれにあたる．この場合，本来名詞1と名詞2が係り得る場合であっても，名詞2と名詞3が複合語を作る場合は，名詞1は名詞3に係ると解析される．また名詞1と名詞3が複合語を作る場合も，名詞1は名詞3に係ると解析される．しかし，名詞3が名詞2とともに複合語を作る場合（「日本大使館」の場合）は，名詞1はヘッドに近い名詞2に係るとみなせる．また，名詞3が名詞1と複合語を作る場合（「LZH ファイル」の場合）も，名詞1と複合語がほぼ同じものを指す場合には，名詞1と名詞2の係りを認める方が自然である．そこで，2つの名詞の間に名詞1語を挟むパタンを加える．これは，例えば「東京　面積」というキーワードに対して「東京ドームの面積」という句を含む文書を拾い上げるなど，失敗することも起こりうるが，我々の実験ではこのパタンを拾い上げる方が全体的な性能は良くなることを確認している．

以上の検討に基づき，本論文では表3.2に示すパタンを抽出して解析対象とする．

名詞以外のキーワードを含む場合

次に，名詞以外の語がキーワードとなる場合について考える．まず，キーワードとして選択されるものを自立語に限定する．自立語には，名詞・代名詞・連体詞・副詞・接続詞・感動詞・動詞・形容詞・形容動詞がある．この中で検索キーワードとしてまず用いられることのない，代名詞・接続詞・感動詞を除き，名詞・連体詞・副詞・動詞・形容詞・形容動詞について組み合わせを考える．

しかし，前述のように2語キーワードのうち少なくとも一方は実体名詞であることがほとんどであるため，実体名詞と他の品詞（連体詞・副詞・動詞・形容詞・形容動詞の5品詞）の語の組み合わせだけに限定する．この内，副詞（及び，形容詞連用形，形容動詞連用形）は連用修飾を構成する語で実体名詞と組になって用いられることは希であるので，候補から名詞と副詞（及び，形容詞連用形，形容動詞連用形）の組を除外する．

残りの候補中で，まず，連体詞・形容（動）詞連体形は，実体名詞とともに用いられる場合はその実体名詞を修飾して何らかの属性値を指定するものと考えることができる．従って，連体詞・形容詞・形容動詞のキーワードが実体名詞のキーワードを直接修飾するパタンを取り出せばよい．また，動詞・形容（動）詞終止形が実体名詞とともに用いられる場合の可能な組み合わせパタンとしては，表3.2で「実体名詞＋現象名詞」の組み合わせで列挙したパタンの内で，(b) と (c) 以外について現象名詞と動詞・形容（動）詞終止形を入れ替えて得られる3つのパタンと，動詞・形容（動）詞終止形が連体修飾節を構成して実体名詞に係るパタン，実体名詞が格助詞を介して動詞・形容（動）詞終止形に係るパタンとなる．

以上を整理し，表3.3に名詞以外のキーワードを含む場合の係り受けパタンを示す．形容詞がキーワードとして用いられる場合，形容詞は終止形と連体形が同じ形であるため両方の可能性を考えて処理する．また，形容動詞が用いられる場合，語幹だけが入力されることがほとんどである．この場合も活用形が判断できないため，終止形・連体形の両方の可能性を考えて処理する．

表 3.2: 名詞キーワード間の係り受けパタン

• 実体名詞＋現象名詞
(a) 実体名詞と現象名詞が連接し，実体名詞が現象名詞に係る．
(a) 現象名詞と実体名詞が連接し，現象名詞が実体名詞に係る．
(b) 実体名詞が連体助詞を介して現象名詞に係る．
(b) 現象名詞が連体助詞を介して実体名詞に係る．
(c) 現象名詞が動詞に係り連体修飾節を構成して実体名詞に係る．
(e) 実体名詞が連体助詞を介してある名詞に係り，現象名詞もその同一の名詞に係る．(I)
(f) 実体名詞がある名詞 1 語を介して現象名詞に係る．(II)
• 実体名詞＋属性名詞
(a) 実体名詞と属性名詞が連接し，実体名詞が属性名詞に係る．
(b) 実体名詞が連体助詞を介して属性名詞に係る．
(c) 実体名詞が動詞に係り連体修飾節を構成して属性名詞に係る．
(d) 実体名詞と属性名詞がそれぞれ格助詞を介して同一の動詞に係る．
(f) 実体名詞がある名詞 1 語を介して属性名詞に係る．(II)
• 実体名詞＋値名詞
(a) 値名詞と実体名詞が連接し，値名詞が実体名詞に係る．
(b) 値名詞が連体助詞を介して実体名詞に係る．
(d) 値名詞と実体名詞がそれぞれ格助詞を介して同一の動詞に係る．
• 実体名詞＋実体名詞
(a) 双方の実体名詞が連接し，一方がもう一方に係る．
(b) 一方の実体名詞が連体助詞を介してもう一方の実体名詞に係る．
(c) 一方の実体名詞が動詞に係り連体修飾節を構成してもう一方の実体名詞に係る．
(d) 双方の実体名詞がそれぞれ格助詞を介して同一の動詞に係る．
(f) 一方の実体名詞がある名詞 1 語を介してもう一方の実体名詞に係る．(II)

表 3.3: 名詞以外のキーワードを含む場合の係り受けパタン

● 連体詞・連体形形容（動）詞＋実体名詞
(a)　連体詞・連体形形容（動）詞が直接実体名詞に係る.
● 動詞・終止形形容（動）詞＋実体名詞
(a)　（格助詞が省略され）実体名詞が直接動詞・終止形形容（動）詞に係る.
(e)　実体名詞が連体助詞を介してある名詞に係り，動詞・終止形形容 (動) 詞が連体修飾節を構成して同一の名詞に係る.
(f)　実体名詞がある名詞に直接係り，それが格助詞を介して動詞・終止形形容 (動) 詞に係る.
(g)　動詞・終止形形容（動）詞が連体修飾節を構成して実体名詞に係る.
(h)　実体名詞が格助詞を介して動詞・終止形形容（動）詞に係る.

3.1.3　文におけるキーワードの重要度

文の主たる主張は通常，主動詞周りやヘッドの名詞で述べられる．主動詞の格名詞に係る連体修飾節は，格名詞の指示対象（referent）を制限することが主な役割である．また，主節動詞に係る従属節は，主節で述べる命題の前提や原因などを述べるものである．従って，該当する係り受けパタンが連体修飾節や連用修飾節内に見つかったとしても，文の主たる主張はその係り受けパタンとは異なる実体や現象について述べられることが多い．体言止めの表現の場合でも，該当する係り受けパタンがヘッドの名詞を修飾する連体修飾節内に見つかった場合は，体言止め表現で焦点を当てている対象とは異なる実体や現象について述べられることが多い．そこで，2 つのキーワードのうちの少なくとも一方が 2.1 節で定義した主題を構成する要素となるように（すなわち，主節の構成要素あるいは体言止めのヘッドの名詞となるように），該当する係り受けパタンの文中における位置を判断材料に加えることを検討する．

基本的には，係り受け関係のある 2 つのキーワードが文に含まれており，少なくともそのうちの 1 語が主文中に存在する場合，もしくは 2 つのキーワードの内の一方が連体止め表現の末尾の名詞となっている場合 (A) と，連用修飾節または連体修飾節内にしか係り受け関係のあるキーワードが存在しない場合 (B) とに分類する．このためには，基本的には構文

解析を施し，該当する係り受けパタンがどこにあるかを判定すればよい．しかし，接続助詞を用いた従属節を多用する文は，全体の語調が長くなり，構文解析で失敗してしまうことが多い．これは，接続助詞が離れた語に係りやすく，係り先を特定しにくいことによる．

そこで，連用修飾節の中に 2 語の係り受けパタンを含む場合の判定は，表層の形態素の順序に基づいて行い，連体修飾節の内部か否かの判定は構文解析結果を用いて判定を行う．具体的なアルゴリズムについては，4.1.2 節及び 4.1 節で詳述する．

3.2　表構造による係り受け関係

3.2.1　表の基本構造

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，分かり易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 [19]．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 [20]，大前ら [21]，吉田ら [22]，板井ら [23] 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み合わ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，それを問い合わ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 [24] 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 [25]．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いる [25]．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良いかが明らかではなかったためであると考えられる．

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持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ンは以下の3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．

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3.3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．

- **基本単位**
  一つの情報（実体，現象あるいはそれらの組み合わせ）を1行で構成する．

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 3.3: タイプ (i)

- **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．補間する表現は全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 3.3）の場合，全ての行について「○○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**
  行毎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関わることが多い．行見出しは，無い場合もある．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 3.3 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．

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 3.4 に示し，タイプ (i) 同様，項目ごとに述べる．

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 3.4: タイプ (ii)

- **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，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無く，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 2 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，行および列毎に付加される「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，同一行およ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．補完する表現は全てのセルに対し共通である．例えば，郵便物の料金表（図 3.4）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）全てのセルについて「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共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 3 つ以上の場合は，見出しの階層構造化が起こる．例え

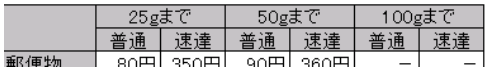

図 3.5: 見出しの階層構造化

ば「速達かどうか」という条件が加わった場合は，図3.5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「条件の組み合わせで何を規定しているのか」を示すとみなせる．図3.4では，郵便物の種別と重さで料金を規定している．

3.2.2　表の係り受け構造

表構造は，3.2.1節で説明したように，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持つ表の構成要素の中の2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2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み合わせから推定される「係り受けパタン」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，3.2.2節の末尾で述べる．）

3.2.2節では，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ンを整理する．次いで3.2.2節で，上述の近似を用いた場合にそれらのパタ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．

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 4 つと捉えたとき，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 2 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，可能な組み合わ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，2 つの検索キーワード間に想定しうる検索意図と関連がある可能性がある場合は，その組み合わせを検討対象とする．

また，表と意味的関係を持ちう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[8]．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

[8]（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．

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み合わ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良い.

(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1つを特定するキーと考えられる．例えば，図3.4における「スクリプト言語」と「Python」の組み合わせであり,「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／列見出しはセル毎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．この組み合わせによる検索は，その属性値を持つ実体／現象についての詳細な情報を求める場合に起こりうる．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 1 つの具体的現象「120 円（かかる)」，という組み合わせであり，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み合わ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」でさら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え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み合わせの場合,「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み合わせの場合，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／現象の他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み合わせではないが，ほとんど関係の無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無い場合は，意味的関係が無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，2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，行見出しに答えがある．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・同列かどうかに関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捉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うることから，係り受けパタ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ンとは言えないため，取り扱うべき組み合わせから除く．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
(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

またさらに，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．

表 3.4: 表構造における係り受けパタン

記号	検索キーワードの出現パタン
(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

以上の議論を整理し，HTML 文書内の表構造における 2 語が出現するパタンについて，何らかの意味的関係にあると思われるパタンを中心に整理を行った（表 3.4）.

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.

適合可能性に基づいた係り受けパタンの検討

前節では，2 つのキーワードが共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，抽出する意味を持ちうるものとして表 3.4 のパタンを挙げた．ここではそれらのパタンについて，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）[9]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 3.5）. 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[10]

③ （①②いずれにも該当しないながら）キーワードがひとつの表内に頻出（暫定的に閾値を 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わかった.（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関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，1 つのレコードが 1 つの名詞句を構成できる場合や 1 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，1 つのレコードでもセルによって別の述語を持つ場合もあり，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言い難い．このため，想定したような検

[9]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.

[10]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.

表 3.5: ウェブページ中の表を目視により分析調査した結果

n=500(ページ)

表の係り受けパタン	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%

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')，さらに文の係り受け判定に依存する (E) を除いたパタンを，表の係り受け判定として採用することにする．

3.3　見出し構造による係り受け関係

3.3.1　見出し構造の利用方針

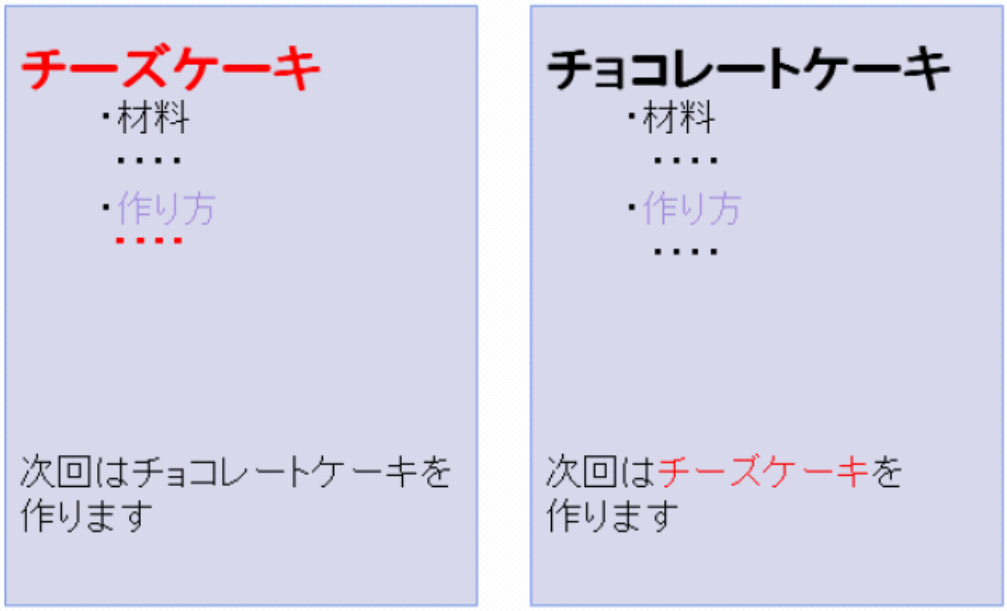

図 3.6: 見出しに含まれる語とページの内容

図 3.6 の 2 つのページはともに「チーズケーキ」と「作り方」という 2 語で検索した結果の例であるが，その内容は全く異なる．左のページはチーズケーキに関して材料や作り方が掲載されていると推測できるが，右のページはチョコレートケーキに関する内容が主であり，チーズケーキに関してはページの末端部分に少し触れられているのみとなっている．

「チーズケーキ　作り方」で検索した場合の検索意図に適っているのが左のページであることは明らかである．この違いは，ウェブページの階層構造が深く関わっていると考えられる．つまり，左のページは上位階層の「チーズケーキ」がページの主体になっていて，かつ，その下位階層に「作り方」が存在している．一方，右のページは中心の話題が「チョコレートケーキ」であり,「作り方」と「チーズケーキ」は無関係な位置に出現している．

すなわち，上位階層に出現する語と，その下位階層に出現する語には，

強い意味的な結びつきがあり，検索意図に適ったページを探すことに利用できると考えられる．

そこで，ウェブページに存在する階層構造に着目する．我々は，階層構造の解析に，見出しの階層関係を利用する．すなわち本システムは，ページ内に存在する見出しとその見出しが影響を及ぼす範囲（以下，支配範囲と呼ぶ）を抽出する．一方のキーワードが見出しに含まれるとき，その見出しの支配範囲内にもう一方のキーワードが含まれていれば，キーワード間に強い意味的関係が存在する，つまり，適合ページであると推定する．

3.3.2 関連研究

ウェブ文書の見出し・箇条書き構造に着目した研究として，単語の上位下位関係を WWW 上の大量の HTML 文書より自動獲得することを目指す新里らによる研究 [26]，HTML 文書から見出しの階層構造を抽出する松本らによる研究 [27] が挙げられる．

新里らは,「HTML 文書中に現れる箇条書きやリストボックス，テーブルのセルなどの『繰り返し』の要素は，意味的に類似しており共通の上位語を持ちやすい」という仮説に基づき,「タグの繰り返し」を抽出することによって箇条書きなどを抽出する．つまり，箇条書きの判断は，タグに基づいている．このため，タグを使わずに視覚的に表現されている箇条書きには対処していない．

松本らは，HTML 文書から階層構造を抽出することを目的としている．HTML 文書をブロックレベル要素で分割したものをノードとして階層化し，そのうち子ノードを持つノードを「見出し」とみなす．すなわち，同じ階層の箇条書き項目同士であっても子ノードを持つ／持たないによって，見出しであったりなかったりする．また，階層の判断の材料として，(1)HTML のタグの階層構造に着目した DOM のパス，(2) 各テキストセグメントのインデント，(3) 文頭記号の相違や有無，テキストセグメントの長さ，句読点の有無，品詞情報などの言語的情報，の 3 つを用いている．しかし，文字の大きさや強調といった視覚的なタグを，階層の判断材料として用いていない．

我々も，ページ全体を（独自の定義による）ツリー構造としてとらえる．しかし，見出しかどうかの判断は，子ノードを持つかどうかには依存しない．子ノードを持たない箇条書きなどの構造についても「支配範

囲を持たない見出し」としてとらえ抽出を行う．また，他の手法では用いられていない階層を判断材料とし，ウェブページを視覚的にとらえることにより得た情報を幾つか利用することにより，見出しによる階層の判断の精度を高める工夫を行う．

3.3.3　ウェブページの構造化と構造間の関係

ウェブページ内には多様な「見出し」が存在し，それぞれの見出しには支配範囲が存在することは，大多数の人の共通認識であると思われる．しかしながら,「これは見出しである」と感じる基準は人によって様々であり，主観に依存する．

そこでまず，本研究における「見出し」を以下に定義する．

[見出しの定義]

(A) 一行の短い文で書かれており，他の見出しや文，図，表に対し一目で内容が分かるように付けられた標題.
(B) 事柄をいくつかに分けて書き並べている一つ一つ，他の見出しや文，図，表の標題にはならないものもある．箇条書き.

図 3.7 において，(A) の「見出し」とは，③⑨などである．また (B) の「見出し」は，⑪などの箇条書きを指す.

続いて，ウェブページの文書構造について，図 3.8 のように整理する．

[Web ページの文書構造]

図 3.7 に示した webPage は pageTitle である①とそれ以降の body 部分である②の 2 つで構成されている．次に，その body の部分②は 1 つ以上の contents より構成される．図 3.7 の例では contents は 1 つで block から構成される．この block は③の headline とそれ以降の④の body から構成されている．この body は 4 つの contents から構成されそのうち 2 つは element である．element の 1 つは⑤で sentence であり，もう 1 つは⑥の figure である．残りの 2 つの contents はともに block であり，1 つは headline に⑨，body に⑩をもつ block であり，もう 1 つは headline に⑦，body に⑧をもつ block である．このようにとらえることにより，空白であるページ以外のすべてのページにおいて，図 3.8 で定義したウェブページの文書構造によって表すことが可能である．

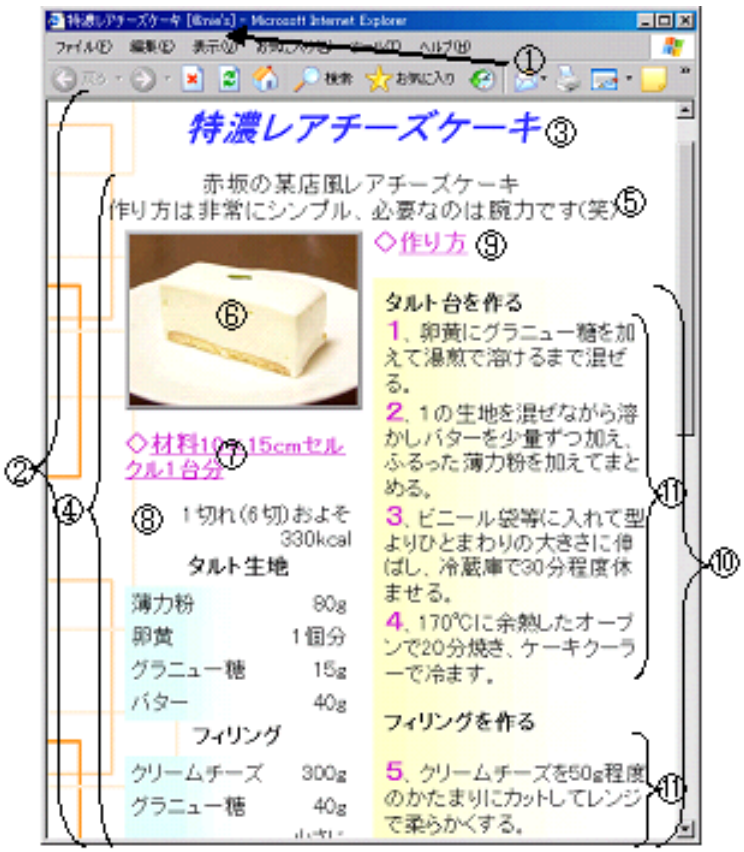

図 3.7: ウェブページの例

1つの webPage は body という塊で階層を構成しており，body は pageTitle や headline の後に現れ，先行する pageTitle や headline のテーマについての記事を構成していることが i と iv から分かる．このことから，pageTitle と headline の階層構造を取り出すことでウェブ文書全体の階層をとらえることができると考えられる．

このようにしてとらえたウェブページの文書構造は，文書内に現れるキーワード間に意味的関係があるか否かを判定する手がかりとなりうる．すなわち，ページタイトルや「見出し」が一方のキーワードを含み，そのボディがもう一方のキーワードを含む場合，それらのキーワード間には文書の階層構造上の上位下位関係，つまり係り受け関係（＝意味的関係）があると言える．

つまり，ウェブページの構造から，以下の (1)〜(8) には係り受け関係があると考えられる．これを「ウェブページの文書構造による係り受け

```
webPage ::= pageTitle body                    ··· i
body    ::= contents +                        ··· ii
contents ::= block | element                  ··· iii
block   ::= headline body | headline          ··· iv
element ::= table | sentence | figure         ··· v
```

図 3.8: ウェブページの文書構造（BNF 表記）

関係」と定義する．

(1)「ページタイトルと見出し」
(2)「ページタイトルと表」
(3)「ページタイトルと地の文」
(4)「ページタイトルと図」
(5)「見出しとそのボディ内にある見出し」
(6)「見出しとそのボディ内にある表」
(7)「見出しとそのボディ内にある地の文」
(8)「見出しとそのボディ内にある図」

(1)〜(8) の中で，本研究の趣旨である「見出し」に関係する部分に着目すると，(1)，(5)，(6)，(7)，(8) に絞られる．このうち (6) の「表」については，3.2 節で述べた「表構造による係り受け関係」と併せて，今後検討を行っていく予定である．また (7) の「見出し」と「地の文」という関係は，現段階では研究対象から外している[11]．(8) の「図」については，alt タグで指定されたテキストについてのみ，取り扱う．よって，本論文においての「見出し構造による係り受け関係」は次のように示せる．

［見出し構造による係り受け関係］

(a)「ページタイトルと見出し」
(b)「見出し間の親子関係」

この 2 つを「見出し構造による係り受け関係」と定義する．

先ほど挙げた (1)〜(8) のうち，(5) は上記定義の (b)，(a) に関しては (1) である．

[11]なぜなら「地の文」は範囲が広く，その範囲におけるキーワードの重要度をつかみにくいからである．また「地の文」は，支配範囲（情報量）を持たないことから，すなわちページ内におけるキーワードの重要度が低い，とも考えられる．そういう意味では，例えば地の文であっても，キーワードがリンクの場合は取り扱うべきかもしれない．

先程の図 3.7 は「チーズケーキ」と「作り方」という 2 語で検索した結果現れたページであるが，ページ全体の標題となる「見出し」である「特濃レアチーズケーキ」に検索キーワードが含まれ，その支配範囲内にあり親子関係にあたる「見出し」である「◇作り方」にもう一方の検索キーワードが含まれる．この場合にはキーワード間に意味的関係があると考えられ，実際,「◇作り方」の支配範囲である⑩部分はユーザの欲しい情報であると考えられる．

この「見出し構造による係り受け関係」を取り出すため,「見出し」はもちろん,「見出し」を標題としボディ部分にあたる「見出し」の係る範囲つまり「見出しの支配範囲」を取り出す必要がある．この「見出し」と「見出しの支配範囲」によって,「見出しの階層構造」をとらえることができる．

第4章　係り受け構造の判定方法

4章では，文・表・見出しの各構造ごとに，構造の抽出方法や係り受けの判定アルゴリズムを詳述したあと，試作システムの全体構成を紹介する．

4.1　文の係り受け構造を用いた判定の方法

4.1.1　文判定のアルゴリズム

文の係り受け構造を用いた判定のアルゴリズムについて，以下に示す．なお，本論文では，2つの検索キーワードをともに含んだ文を「キーセンテンス」と呼ぶ．

(1) 各種タグが含まれるHTML文書内から文章を切り出すため，改行コードを取り除き，句点やピリオド，構造の終端を表すタグで文章を切り分ける．但し，ピリオドの前後が数字であった場合は，小数点とみなし，そこでは分けない．また，括弧内の句点においても切り分けない．

(2) (1)で切り出した文章からキーセンテンスを抽出する．

(3) (2)で抽出したキーセンテンスに対し，括弧の記号に応じた整形を行う．(詳細は4.1.2節参照)

(4) キーセンテンスに対し，形態素解析の結果から接続助詞候補の語を探し，存在した場合は，接続助詞かどうかを判定する．接続助詞と判断した場合は文を分割し，接続助詞直前にあった語の活用を終止形に戻す．分割後の文は，いくつに分割されたうちの何文目かという情報を保持し，この情報は(8)で利用する．(詳細は4.1.2節参照)

(5) キーセンテンスに対し，単語を列挙した形態の文かどうかを判定し，該当した場合はすべての読点，カンマ，中黒点でセンテンスを分割する．(詳細は4.1.2節参照)

(6) 自然言語処理パーザによって分割後のキーセンテンスを構文解析する.

(7) 構文解析されたキーセンテンスと，検索キーワードのカテゴリによって用意された係り受けパタンを比較する．キーセンテンスの一部がこれらのパタンのいずれかと一致した場合に，入力キーワード間に修飾－被修飾関係があると判定する.

(8) (7) によって修飾－被修飾関係と判定された場合は，(4) の結果に基づき，元々連用修飾節内にあった文かどうかを確認する．これに該当する場合は，係り受け関係のあるキーワードのうちヘッド（文末）に近い方の語が主格・提題格を構成し，なおかつ格助詞が「は」であった場合を除き，そのキーセンテンスは (7) で一致しなかった文と同等に扱う．格助詞「は」で取り立てられている場合は，係り受けパタンと一致したとみなす．一方，(4) で分割された文の中で一番末尾の文と係り受けパタンが一致した場合は，その係りの位置が，連体修飾節内か否かを構文解析の結果から判定し，連体修飾節内のみしかパタンに一致していなければ，(7) で一致しなかった文と同等に扱う.（3.1.3 節参照）

(9) (7) においてパタンと一致しなかった場合は，並列構造解析を行う.（詳細は 4.1.2 節参照）

(10) 文中に「<」が無く「>」が存在する場合，キーセンテンスがカテゴリ階層を表しているとみなし,「>」を挟んでキーワードが出現していても修飾－被修飾関係があると判定する.（詳細は 4.1.2 節参照）

4.1.2 係り受け解析の工夫

3.1.2 節および 3.1.3 節の分析に基づいてシステムを実装するにあたり，ダウンロードした検索対象ファイルから HTML タグを除去し，できるだけ正確に文章を切り出す必要がある．しかし，これだけでは前処理として不十分である．ウェブページは，新聞記事など一般の文書より，括弧などの記号・口語・誤字・辞書にない語などが多いため，構文解析精度が低くなる傾向にあるという報告がなされている [28, 29, 30]．また，長文や単語が列挙されているような並列構造も，解析誤りの原因となりうる.

実際に，様々な 2 語検索の結果ページから検索キーワード同士が係り受け関係にある（と人間が判断できる）文をページ毎に 1 文ずつ計 100 文集めて構文解析を行い，2 語が係り受け関係にあることが正確に判断できているかどうかを調査した．この実験の文の係り受け解析には，我々がシステムに導入している（株）CSK で開発された日本語パーザと，CaboCha [31] を利用した．実験の結果，精度はそれぞれ 75%，77%と低いことがわかった．本来 CaboCha の係り受け正解率[1]は 89.29%とのことであるので，やはり精度は下がっている．ゆえに，パーサの誤解析を減少させるための工夫を行う必要がある．

本論文では以下に述べる方法で対処を行う．この工夫の効果を定量的に評価するためには本来，大規模なデータに基づいて検証すべきであるが，本論文では係り受け解析そのものが主目的ではないので 100 文程度の予備実験でとどめた．この 100 文の範囲内においてではあるが，解析精度が（CSK パーザにおいて）85%に向上することを確認した．

括弧の処理

実際のデータを調査した結果，主要な括弧の用途として，①強調・引用，②語や節の補足説明，③見出しなどを表現するための装飾，の 3 つのタイプが存在することがわかった．これらについては用途ごとに用いられる記号も異なっており，多くの場合，①にはカギ括弧（「」『』），②には小括弧，③の場合はその他の括弧（【】［］＜＞など，以下便宜上“ 装飾括弧 ”と呼ぶ）が良く用いられる傾向にある（表 4.1）．

そこで括弧の記号に応じた整形を検討する．いずれの括弧の場合でも，一対の括弧内にキーワードが両方とも存在する場合は，括弧の中だけを解析対象とする．ともに括弧外にある場合や，一方のキーワードのみが括弧内にある場合は，カギ括弧の場合は括弧のみを削除し，それ以外の括弧の場合は括弧内のフレーズごと削除する．小括弧の場合はこのようなケースも正しく構文解析されるべきだが，そのためには括弧の中と外の意味を照らし合わせて，括弧内のフレーズが括弧外のどの語・句・節・文をどのように補足しているかを判断する必要がある．これについては現状では対処できないため，括弧および括弧内のフレーズを削除することにしている．これは，そうすることで構文解析の精度をあげる方が全体として精度向上につながるためである．装飾括弧の場合は，括弧内の

[1]文末の一文節を除くすべての文節に対して, 正しく係り先が同定できたものの割合

表 4.1: 用途ごとの括弧記号の使用率

調査対象；ウェブページから抽出した，括弧のペアを少なくとも 1 つ以上含む 1190 文

n=1327(括弧ペアの数)

用途	カギ括弧	小括弧	装飾括弧	計
①	29.0%	0.1%	0.4%	29.5%
②	0.0%	49.9%	1.7%	51.5%
③	0.7%	0.2%	11.2%	12.1%
※	0.0%	6.8%	0.2%	6.9%
計	29.7%	57.0%	13.3%	100.0%

※…番号付リスト，曜日，注釈・図表の参照,（株）（代）などの省略，顔文字，数式，など

フレーズが括弧外と構文的な繋がりをもたないことが多い（我々の調査ではこれに該当しないケースはいずれの括弧とも 5%未満であり，無視できると判断）．装飾括弧において括弧内外に構文的な繋がりはなくとも意味的関係は存在する場合（例；【大阪】ホテル）がありうるが，それは文構造ではなく見出し構造を用いた修飾－被修飾関係の表現とみなすべきであり，本論文では扱わない．

長文の分割

一般に，構文解析の精度は長文に対しては大きく低下する傾向にある．本論文で提案する手法では，文全体の構文木を求める必要はなく，キーワードとして指定された 2 語が修飾－被修飾関係をもって出現するか否か，および，その文中における大まかな位置（主文に少なくとも 1 語が出現するか否か）を求めればよい．そこで，長文については以下の方法で分割を行い，構文解析の精度の低下を押さえる．2 語のキーワードがそれぞれ従属節と主節に分かれて存在する場合，それらの 2 語間に係りはないとしてよい．従って，従属節を主節と切り離して解析することとする．そこで，形態素解析後，構文解析を開始する前に，従属節の分離処理を行う．そのために，まず，接続助詞を探す．例えば「が」「から」のように表層は同じであるが接続助詞とさらに別の品詞をもつ語（多品詞語）の場合，接続助詞の直前につきうる品詞は述語，助動詞，終助詞類に限

定されることを利用し，接続助詞か否かを見分ける．接続助詞が見つかると，文を一度切り離し，接続助詞直前にあった語の活用を終止形に戻す．そして分割後の文に対してそれぞれ構文解析を行い，係り受けパタンとの比較処理を行う．接続助詞より文頭側の文に，係り受けパタンに一致する箇所があった場合は，一致した部分が提題の副助詞「は」で取り立てられた場合を除き，適合とは判定しない．

列挙された単語の分割

ウェブページには，単語が数多く列挙されていることがよくある．例えば，ホテルの予約を行うサイトにおいて，検索され易くすることを目的として,「ホテル予約・宿泊予約・格安ホテル」などと関連ワードをページ中に列挙しておくケースなどが挙げられる．この場合，列挙された個々の語が複合語であり，その中に係り受け構造が存在する場合もある（「ホテル予約」＝「ホテルの予約」,「格安ホテル」＝「格安のホテル」など）．そのような係り受けの解析をしておく必要があるが，このままの形でパーザにかけると，パーザの解析誤りの原因となる．そこで，このような単語の列挙については，文中のすべての読点，カンマ，中黒点（但し括弧内を除く）でキーセンテンスを分割し，分割後のセグメントも文とみなし，各々に対し構文解析を行う．

そのためにはまず，自立語が列挙された構造なのか否かを判別しなければならない．列挙構造の場合，基本的に文を構成するのは自立語のみと考えられる．但し列挙の一成分の中に連体助詞「の」が入り込むことはありうる．そこで，センテンス中に「の」以外の付属語（助動詞・助詞）がひとつも存在しないかどうかをチェックする．続いて，文中の（括弧内を除く）すべての読点，カンマ，中黒点で分割したとき，分割後の文のいずれかがキーワードを2語とも含んでいることを確認する．キーワードを2語とも含む文が存在しない場合は，分割せずにそのまま構文解析や次項で述べる並列構造解析を行うことによって修飾－被修飾関係が正しくとれる可能性があるので，分割は行わない．

以上の手順で列挙された単語の分割を行うことによって，例えば，検索キーワードが「マレーシア　電圧」のとき，キーセンテンスが「マレーシアの電圧・周波数・プラグ.」であった場合は「・」で分割され「マレーシアの電圧.」となる．

並列構造解析

単語が並列の関係で列挙されている場合，同一の構文であっても複数の解釈が可能なため，係り受け関係の抽出は困難であり，我々が利用したパーザでは並列関係を正しく抽出することはできない．そこで，パーザの弱点を補うため，典型的な頻出するパタンに対して並列構造の解析処理を独自に行う．ここで典型的な並列構造とは，以下のタイプを指す．なお，Nは名詞，Vは動詞またはサ変名詞,「・」はセパレータを表している．セパレータとは，並列連体助詞及び記号（中黒点，カンマ，読点，アンパサンド）と定める．

- N_1・N_2・$N_3 \cdots N_{m-1}$連体助詞 N_m
- N_1連体助詞 N_2・$N_3 \cdots N_m$
- N_1・N_2・$N_3 \cdots N_m$格助詞 V_1
- N_1格助詞 V_1・V_2・$V_3 \cdots V_m$

すなわち，連体助詞（または格助詞）の直前または直後に一方のキーワードが存在し，連体助詞（格助詞）を挟んで反対側にセパレータと自立語が交互に出現するケースを想定する．もう一方のキーワードが，その自立語とセパレータが交互に並ぶ範囲内に存在していれば，2つのキーワードは修飾－被修飾関係に該当するとみなす．

以上の処理によって，例えば，検索キーワードが「マレーシア　電圧」であるとき,「マレーシアの周波数・プラグ・電圧の一覧」や「マレーシアとシンガポールの電圧」というキーセンテンスが修飾－被修飾関係に該当すると判断することができるようになる．

なお，並列構造解析には以下の問題点が残されている．例文中の〔〕で囲まれた語は検索キーワードを表すものであり，実際のセンテンスに含まれてはいない．

例）〔ＬＺＨ〕ファイルのダウンロードと〔解凍〕。
→ 列挙のルールの中に間に一語が存在することを許可するルールは現状ではない

例）海外旅行、世界の外貨の〔為替〕、通貨〔レート〕情報。
→「通貨」と〔レート〕の間に連体助詞がないので，列挙のルールに適合しない

例）〔世界〕で使用されている周波数・〔電圧〕・電源プラグです。

→ 列挙のルールの中に関係節を介して係ることを許可するルールは現状ではない

例）〔格安航空券〕、ホテルクーポン、パックツアーの〔販売〕。

→ 複合語なので，列挙のルールに適合しない

例）〔ＬＺＨ〕・ＺＩＰ・ＣＡＢ・ＲＡＲ・ＮＯＡ・７７７・ＧＣＡ等の圧縮・〔解凍〕ができます。

→ 一文中に２つの列挙を認めるルールがまだない

階層分類表記判定

ウェブページではページ内容の階層分類構造を表現するのに「＞」を用いることがよくある．例えば,「世界の電圧＞東南アジア＞マレーシア」といった表記は，文とは言い難いが，ここで「マレーシア」「電圧」間には何らかの意味的な関係があると推定できる．

そこで，文中に「＜」が無いにも関わらず「＞」が存在する場合,「＞」を挟んでキーワードが出現していても修飾－被修飾関係があると判定する．

4.2　表の係り受け構造を用いた判定の方法

4.2.1　表判定のアルゴリズム

表判定ユニットの処理の流れを図 4.1 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3.4のパタン（(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無くなるまで，(3)(4)(5) を繰り返す．

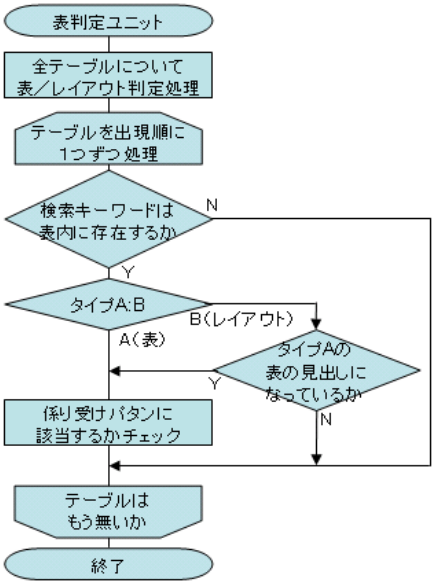

図 4.1: 表判定ユニット

4.2.2 表の抽出

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 [32] 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（図 4.2）は，同じ行や列のセル同士に意味的な深い関係を持たないどころか，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「意味的に表を表わすテーブル（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

Ⓐ の Web ページの中に「表」は存在していないが，隠されたテーブルタグを可視化すると Ⓑ のように存在している．このように，レイアウトを制御することを目的としてテーブルタグは頻繁に用いられている．

Ⓐ

Ⓑ

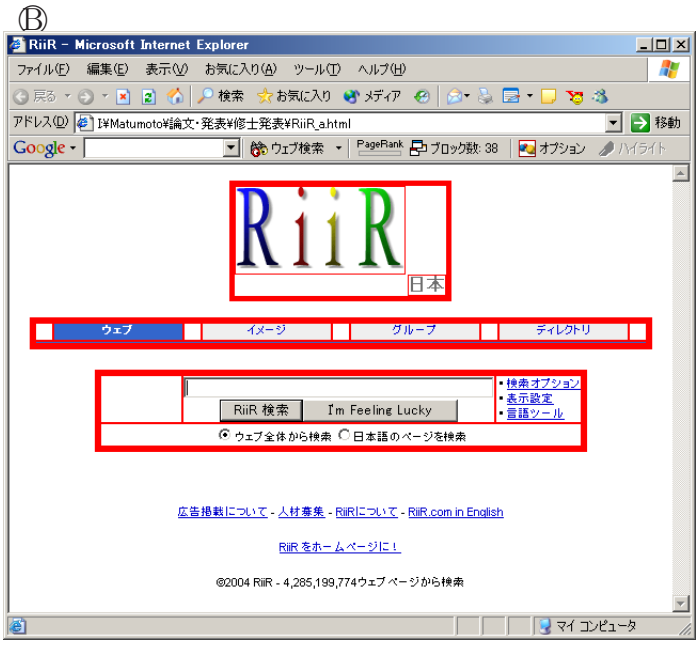

図 4.2: レイアウト制御を目的としたテーブルの例

表 4.2: 表 (タイプ A) ／レイアウト (タイプ B) 戦略

独立変数	①	②	③	④	⑤	⑥	⑦	⑧	判定
	有	有	無	無	＞ 2	!=1	!=1	偽	表
	有	有	無	有	≧ 10	!=1	!=1	偽	表
	有	有	有	無	≧ 10	!=1	!=1	偽	表
※	無	有	＊	＊	≦ 2	＝ 1	＝ 1	偽	レイアウト
※以外	無	有	＊	＊	＊	＊	＊	偽	表
以上以外									レイアウト

＊ … 該当する独立変数の条件式を用いない.

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,

① 線の幅指定の有無
② タグ以外のテキストの存在の有無
③ 句点の存在の有無
④ 色の指定の有無
⑤ セルの数
⑥ リンクのあるセルの数
⑦ 文字のあるセルの数
⑧ テーブル内に 1 行 1 列しかないかどうか

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 [2]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（表 4.2）. オープンテスト（対象：60 ページ，657 テーブル）の結果は，表 4.3 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.

[2]http://www.cse.unsw.edu.au/~quinlan/

表 4.3: 表 (タイプ A) ／レイアウト (タイプ B) 判定結果
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%

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持たないタイプ 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 2 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持つ場合も在り得る．このような複雑な表の扱いについては，田仲ら [33] や大西ら [34] による研究を参考に，別途検討していきたい．

4.3　見出しの係り受け構造を用いた判定の方法

見出し判定ユニットでは，2 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先祖－子孫関係があるか否かを調べる．

4.3.1 見出し構造の抽出アルゴリズム

「見出し構造による係り受け関係」を検索に利用するためには,「見出し」とその「階層構造」を，HTML 文書から正確に抽出する必要がある.

見出しの抽出手法

HTML を用いて表されているページにはタグ本来の見出しや箇条書きの機能を用いて「見出し」を表しているものと，レイアウト機能を用いて表しているものが存在する.

(1) タグ本来の機能を用いて表している「見出し」
- 見出しタグ，キャプションタグ，用語定義タグ，テーブル見出しタグ，リストタグやリストアイテムタグを用いているものは「見出し」として抽出する.

(2) レイアウト機能を用いて表している「見出し」
- 記号・画像（50 × 50pixel 以下）または数字や文字が文頭に連番で存在し，改行タグや段落構成タグが来るまでを「見出し」として抽出する.
- 一行全体や一列全体に強調タグを用いたものや文字に色が付いているもの，リンクとなっているもの，装飾用括弧 [35] を用いて表しているものは「見出し」として抽出する．ここで装飾用括弧とは，隅括弧や大括弧など，見出しなどを表現するためによく用いられる装飾目的の括弧のことである.
- 1 つのセルの終わり部分にコロン記号がついているとき，セル内コロン以前が「見出し」でコロン後を支配していると予測できるので,「見出し」として抽出する.
- 1 つのテーブル内に 1 つのセルしか存在しないとき（複数個のセルが 1 つに結合されている場合も含む），そのセル内は「見出し」として抽出する.

(3) クラス名で指定されている「見出し」
- ページ作成者がスタイルシートにより，クラス名に"TITLE"や"MIDASHI"を用いた場合，それらは「見出し」を意味していると予想でき,「見出し」として抽出する.

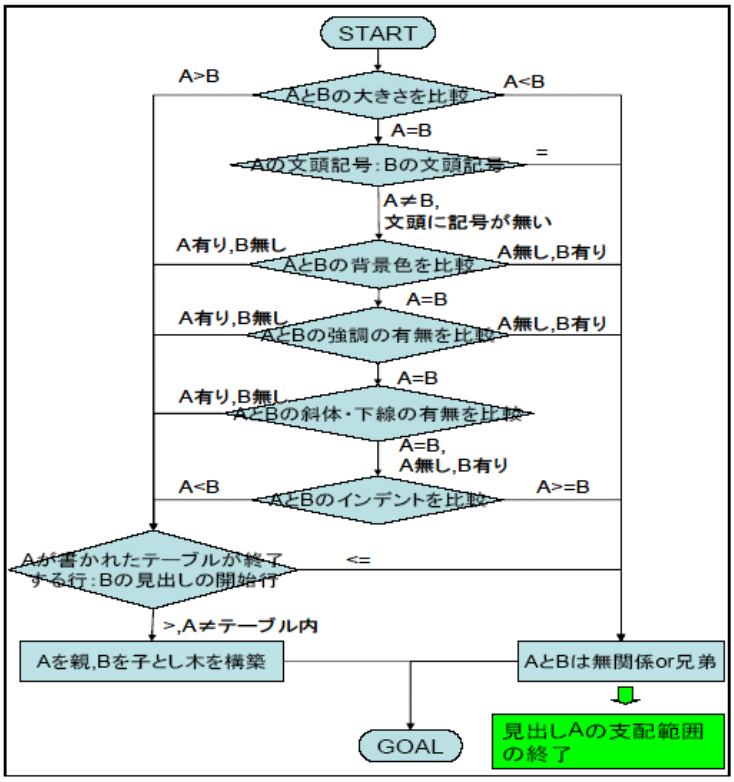

図 4.3: 見出しの階層構造判定アルゴリズム

階層構造の抽出手法

前述の方法で取り出した「見出し」を用い,「見出しによる階層構造」を取り出すアルゴリズムを考案する.

見出しの階層構造判定の材料として,(1) 文字の大きさ（FONT タグ・Hn タグ）による見出しの強弱,(2) 強調タグの有無,(3) 背景色による装飾の有無,(4) 斜体や下線による装飾の有無,(5) インデントによる見出しの文頭の位置,(6)「見出し」文頭の装飾の種類（同じ記号なら同階層）が考えられる. それに加え, 予備実験の結果により判明した,(7) テーブル内に現れる「見出し」の支配範囲はそのテーブル内でありテーブルの外の「見出し」は支配しない, という特徴を含め, 今回これら (1)〜(7) を用いてヒューリスティックに基づき決定木（図 4.3）を作成した. これを「見出しの階層構造判定アルゴリズム」とする. 図中の A は支配範囲を特定中の見出し, B は A の次に出現する見出しである. このプロ

グラムは，ページ中から見出し抽出手法で取得した「見出し」を比較し，それらをノードとした木を作成する．この木により見出しの支配範囲を得ることができ，その情報量も判定できる．図 4.4 は図 3.7 のページから「見出しの階層構造判定アルゴリズム」で作成した木の一部分である．

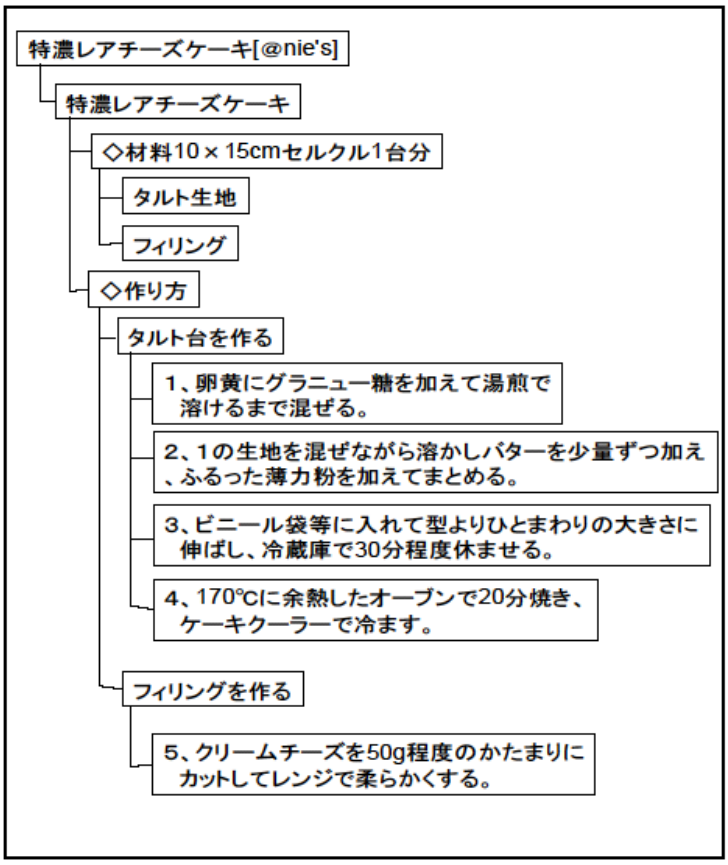

図 4.4: 見出しをノードとした木（一部抜粋）

見出し構造による係り受け関係の抽出

提案手法によって作成した「見出し」をノードとした木を利用し，ページ内の「見出し構造による係り受け関係」の有無を取得できると考えられる．つまり，木を利用しキーワード間に係り受け関係があるか否かの判定を行う．

例えば，2 語の検索キーワード「チーズケーキ　作り方」で，図 3.7 に示すページが出現したとする．このページに対して,「見出し」を抽出し,「見出しの階層構造判定アルゴリズム」を実行した結果，図 4.4 のような「見出し」をノードとした木を生成したとする．

この木に着目すると，検索キーワード「チーズケーキ」と「作り方」を含む「見出し」間の関係は先祖－子孫関係となっており，そのことから，このページにおいてキーワード間には「見出し構造による係り受け関係」が存在すると言える．

この，キーワードを含む「見出し」の先祖－子孫関係の有無により検索意図に適ったページか否かの判定を「見出し判定」とする．

4.4　試作システムの構成

本システムは，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[3]を用いており，システム全体については Ruby で実装している．

システム全体の流れを図 4.5 に示す．

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．

2. 各キーワードを独自の概念階層辞書を用いて実体，現象，属性，値の 4 つのカテゴリのいずれかに分ける．

3. 各ウェブページは文構造，表構造，見出し構造のそれぞれについて係り受けパタンにマッチするか否かを判定される（それぞれのユニットを表判定，文判定，見出し判定と呼ぶ）．

4. 各判定ユニットで，パタンにマッチしたページをキーワード間の意味関係が強いとみなし，一定のスコア（「1 つ以上のパタンにマッチしたら 1，1 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 ソートキーを元の検索エンジンの順位とする手法でランキングを行う．

[3]Google SOAP Search API http://www.google.com/apis/

5. こうして並び替えられた検索結果はユーザに提示される.

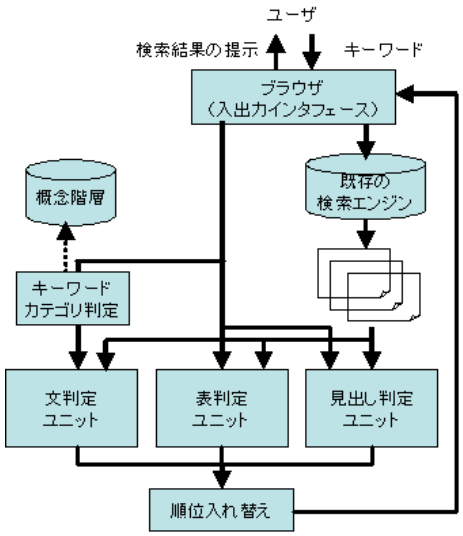

図 4.5: システム全体図

第5章　評価実験

5章では，まずウェブ検索における評価手法を紹介し，我々が行った評価手法について述べる．そしてそれに基づいて行った評価実験の結果について説明し，考察を行う．

5.1　評価手法

5.1.1　従来の情報検索とWeb検索の評価法

検索が成功したかどうかを評価するための伝統的な指標は，精度（precision）と再現率（recall）である．精度や再現率は，適合度による順位付けを行わない伝統的なブール型検索に適した評価手法である．しかし，Webページの検索は ① 全文検索， ② 膨大なページ数，という2つの理由から，検索結果が数多く出力されてしまう [36] ため，適合度順に順位付け出力が行われる．従って，Web検索を評価するには，いくつかの工夫を加える必要がある．適合度順出力の評価のためにTREC(Text REtrieval Conference) [1]を中心とした検索実験で使用されている主要な指標には，精度（λ），再現率（λ），再現率50%での精度，R精度，平均精度などがある [1].

またWeb検索において，ユーザは上位の結果しか参照しない傾向があるため，一般的に再現率（取りこぼしが少ないこと）より精度（不適合ページが少ないこと）が重視されている．

5.1.2　テストコレクションの作成

我々のシステムを実験・評価するためには，テストコレクション，すなわち

[1]TREC: http://trec.nist.gov/

(A) 検索対象となる文書集合（データベース）
(B) 検索質問の集合
(C) 各検索質問に対して各文書が適合しているかどうかの情報
が必要である．(C) のデータセットを構築する作業は容易ではない．なぜなら人手によって全てのページをひとつずつ判定しなければならないからである．そこで広大かつ多様な Web 空間においてどのようなデータ収集を行えば，我々に可能なレベル（コスト）で説得力のある評価データを揃えたい，という欲求を満たせるかを検討した．

そこで，以下の要領でオープンテスト用のデータセットを新たに作成した．手順は以下の通りである．

◎ 実験用データ案 100 組＋α（予備）の作成手順

1. Yahoo!のトップページから幅優先で 6 階層分の URL リストを作成
   →約 20 万件（重複除く,jp サイトのみ）
   理由）検索エンジンのインデックスを作成するために Web ページを収集するロボット（クローラ) は，通常 Yahoo!などのメジャーなポータルサイトからリンクをたどっていく方法で収集が行われるため
2. 1. からランダムに 10 万件の Web ページのダウンロード（1 ページ／秒）
3. バイナリファイル判定・タグ除去・漢字コード変換
4. 形態素解析（0.5 ページ／秒）
5. ページ毎に語の出現頻度をカウント → TF-IDF の計算
6. 2 語の関連が強いものを抽出
   （1） TF-IDF 上位 400 語を抽出，そこから検索ーワードとして不適当なもの（73 語）を除く
   ＜不適当なもの＞
   - 上，新，全，内など，状態を表す一文字，計 12 語
     （理由）もう一語を追加しても検索意図が不明瞭である
   - 歳，件，回など，単位を表す一文字，計 9 語
     （理由）属性値を検索する場合，単位より属性を検索キーワードとする方が自然
   - 数字，曜日を表す一文字，計 9 語
     （理由）もう一語を追加しても検索意図が不明瞭である
   - 収集ページの偏りによる専門用語（個別銘柄，純資産倍率など），計 35 語
     （理由）Web 空間全体で存在数が多いとは思えない

- 情報提供企業（毎日新聞，ロイター，テクノバーン，フィスコなど），計8語
  （理由）検索者が情報提供企業が提供する情報を知りたいとき，その企業名は通常キーワードにしない

(2) TF-IDF401位〜1000位までの中から実体名詞と現象名詞を抽出し，中でも検索キーワードとしてよく用いられる語を経験的に判断し(1)で減らした分（73語）追加する

(3) 400語の組み合わせ（８万通り）についてGoogleで検索を行い，検索結果数を取得して検索結果数が10万〜500万であったものを抽出する（約9割）
理由） 多すぎるものは2つの語がともに一般的によく使われる語であり検索意図が曖昧すぎ（例：「必要＋できる」），少なすぎるものは，2語の関連性が低いとみなすことができるため

(4) (3)から検索意図が推定できそうなものを大雑把に絞込み，そこからランダムに2000組を抽出する
その際，効率的に抽出を行うため，語のタイプごとにグルーピングし，グループごとに判断を行っていく
（例）グループ「メディア」→CD，DVD，ビデオ，ゲーム
理由） 同じタイプの語は組み合わせることができる語も似ている傾向にあるため
（例）グループ「メディア」なら「発売日」「アイドル」「アニメ」「予約」など

(5) 研究協力者3人で，(4）から検索キーワードが重ならないように，検索意図が明確な100組＋予備（50組程度）を決める

(6) 150組の検索意図が，我々の考える「Webの検索意図7つのパタン」を満遍なくカバーしているかどうかを確認する

(7) ①検索キーワードのタイプについて前回の検索実験と比較する
→ 現在使用中のデータセットと比較実験を行うことができるよう，キーワードパタンが同じバランスで存在するようにキーワードの差し替えを行う
②著者が所属する研究室にて収集した検索キーワードとの語の分布を比較する
→ 実際の検索と比較し，候補キーワード対は固有名詞が少ないことが判明

固有名詞に差し替えることで検索意図が明確になる一般名詞について，固有名詞の割合が実際の検索と同程度になるようキーワードの差し替えを行う
尚，①のキーワードの差し替えにあたり TF-IDF 上位 1000 位まで範囲を広げてキーワードを探すこと，また語の連結（複合語の作成）も認める
②のキーワードの差し替えについては，差し替える前の一般名詞の代表と言える知名度の高いもの，かつ流行り廃りの変動が激しくないと思われるものをデータ作成者が決める

（8） (5) とは異なる第三者 3 名によって個別に検索意図の書き出しを行い，その後突合せを行う
→ 3 人の検索意図が一致しないキーワードについては差し替えを行う（その際 (8) で調整した語のバランスを崩さないように留意する）

（9） 適合・不適合判定基準を作成する

（10） 検索結果の確認を行う（適合ページの量が多すぎるものまたは少なすぎるものを排除）
150 語→ 100 語（以上）を決定

以上の手順で決定した 118 組の 2 語キーワード（実体名詞同士の組み合わせが 63 組，実体名詞と現象名詞が 28 組，実体名詞と属性名詞が 21 組，実体名詞と値名詞が 6 組）を，(B) とした．それぞれのキーワード対について Google を用いて検索を行い，検索結果上位 100 件を取得し，評価用のデータ（A) として用いることにした．検索結果として得られた（ダウンロードに失敗したページ及びバイナリファイルを除いた）11776 ページについて，適合／不適合の判定を行い，(C) を構築した．判定は，3 名の研究協力者によって行い，その判定基準は，各キーワードから想定可能な意図に照らしてそのページが出てくることに対し納得できるかどうか，とした．なお，リンク先を実際に確認しなくても（するまでもなく）検索意図に適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．

5.1.3 評価尺度

比較の評価指標には，100 件の内で適合と判定された集合に対する精度・上位 20 件中の適合ページ数・MAP を用いる．精度は，4.4 節で述べ

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，4.4 節のスコアリングに基づいて順位の並び替えを行い算出する．MAP(Mean Average Precision) とは，検索課題ごとの平均精度の平均であり，ランキングの良し悪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，

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

とする．このとき，平均精度 v は，次式で求められる．

$$v = \frac{1}{R} \sum_{i=1}^{n} \frac{z_i}{i} \left(1 + \sum_{k=1}^{i-1} z_k \right) \tag{5.1}$$

MAP の元になる平均精度は，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 [1]．そこで有意水準 $\alpha = 0.05$ で t-検定を行う．

5.2　文構造についての評価実験

5.2.1　係り受けパタンの妥当性の検証

表 3.2・表 3.3 に挙げた係り受けパタンの妥当性を検証する．この検証にあたり，5.1.2 節で紹介した 118 組のキーワード対からなるデータセットを利用する．11776 ページから，抽出されたキーセンテンス計 12536 文について，キーワード間の係り受け関係の調査を行う．結果を表 5.1 に示す．表中の列見出しの「A」は A が B を修飾する,「B」は B が A を修飾する,「AB」はある語に対し A と B が修飾する，ということをそれぞれ表している．

網掛けのセルは本手法には存在しないルールである．これに着目するとまず，網掛けの部分に該当しているセンテンスの絶対数が少ないことがわかる．精度については必ずしも低くないものも存在するが，サンプル数が少ないため今後データを増やすなどして詳細に検討する必要がある．今回の結果を分析したところ，文書中に複数のキーセンテンスが存

表 5.1: 係りのタイプ別精度

n=12536(キーセンテンスの数)

	(a)		(b)		(c)		(d)	(e)	(f)		(g)	(h)
	A	B	A	B	A	B	AB	AB	A	B	B	A
実体 (A) ＋現象 (B)												
適合	1836	106	104	4	2	1	20	40	156	30	9	110
不適合	287	59	34	1	3	1	14	6	81	8	1	36
精度	0.865	0.642	0.754	0.800	0.400	0.500	0.588	0.870	0.658	0.789	0.900	0.753
実体 (A) ＋属性 (B)												
適合	461	4	128	4	26	2	15	0	103	14	0	2
不適合	196	4	37	0	0	1	5	0	62	9	0	0
精度	0.702	0.500	0.776	1.000	1.000	0.667	0.750		0.624	0.609		1.000
実体 (A) ＋値 (B)												
適合	7	52	0	3	0	0	1	0	2	8	0	0
不適合	2	88	3	17	6	2	28	0	8	4	0	0
精度	0.778	0.371	0.000	0.150	0.000	0.000	0.034		0.200	0.667		
実体＋実体												
適合	2203		310		91		222	0	552		0	0
不適合	1226		82		19		73	0	155		0	2
精度	0.642		0.791		0.827		0.753		0.781			0.000

在しており，別のキーセンテンスがその文書を適合たらしめる要因となっているものが 33.8%（23/68 ページ）あること,（提案手法に含まれていない網掛けのルールにおいて）適合ページであるものは特定のキーワードに集中していることなどが確認されている.

5.2.2 類似手法との対比

キーワードの意味分類から係り受けの制約を行う本手法に対し，以下の観点から比較評価を行う.

- 情報検索における，一般的な係り受け関係を用いた手法との比較
- キーワードの意味分類を行わず品詞から想定されうる可能なパタンを用いた手法との比較

表 5.2: 抽出したキーワード対

		検索キーワード
(1)	実体 実体	コンピュータ　雑誌
(2)	実体 実体	動物　写真
(3)	実体 実体	花　美術館
(4)	実体 実体	京都　庭園
(5)	実体 実体	中国　映画監督
(6)	実体 実体	タレント　日記
(7)	実体 現象	ウイルス　対策
(8)	実体 現象	年金　解説
(9)	実体 現象	カレンダー　ダウンロード
(10)	実体 現象	家電　ショッピング
(11)	実体 現象	曲　検索
(12)	実体 現象	ホームページ　作成
(13)	実体 属性	マスカラ　使い方
(14)	実体 属性	ドメイン　登録料
(15)	実体 属性	イラク　言語
(16)	実体 属性	SMAP　プロフィール
(17)	実体 属性	山口　名物
(18)	値 実体	月間　天気
(19)	値 実体	2000 年　重大ニュース
(20)	値 実体	女性　政治家

評価データ

データセットは 5.1.2 節のデータセットから 20 組（表 5.2）を抽出し，これを利用する．20 組の選定にあたっては，実体＋値の組み合わせは他の組み合わせと比較して，現実の検索で用いられる可能性が低く，また実体＋属性の組み合わせも実体＋実体や実体＋現象より少ない傾向にあることも過去の実験の結果 [37] において分かっているので，これを加味し，カテゴリごとのデータ数に差をつけた．

表 5.3: 提案手法と類似手法の精度・再現率・F 値・MAP

n=1995(ページ数)

	精度	再現率	F 値	MAP
(i)	0.761	0.470	0.581	0.729
(ii)	0.766	0.534	0.629	0.736
(iii)	0.771	0.559	0.648	0.738
(iv)	0.757	0.542	0.632	0.733
(v)	0.724	0.628	0.673	0.733

各手法との比較

以下に挙げた (i)〜(v) の各手法においての実験結果を表 5.3 に示す．なお本節では，係りの使い方による差の検証を行うため，提案手法に取り入れている「係り受けパタンの文中における位置」の戦略は，含めずに判定を行っている．

(i) キーワードが直接係り受け関係をもつ文を含む文書を優先させる．

(ii) 表 3.2 のルールを適用して該当している文を含む文書を優先させる．

(iii) 表 3.2 のルールに並列構造解析を加え，該当している文を含む文書を優先させる．

(iv) (a)〜(h) のルールを（実体・属性・現象・値の区別なく）すべて適用して，該当している文を含む文書を優先させる．

(v) キーセンテンスを含む文書を優先させる．

情報検索における，一般的な係り受け関係を用いた手法として，単純に構文解析結果からキーワード同士が直接係り受け関係にあるものだけを拾い上げる方法がある．表 5.3 の (i) がそれに該当する．この方法は係り受けの制約が最も厳しいので，当然再現率が低くなる．反対に (v) は，2 語が同一文中にありさえすればよく，係り受けの制約が最もゆるいので再現率は高くなるが，精度は下がるはずである．実際表 5.3 に示すように再現率は (i) と比べ約 0.16 上昇し，精度も約 0.04 低下している．しかし精

度の低下は予想よりも小幅なものであった．この原因については，5.2.4節で考察する．

これに対して本手法の細やかなルールを適用する (iii) は，(i) から精度を上げつつ再現率を大幅に上げることに成功している[2]．

続いて (iv) は，表3.2・表3.3に示したルール (a)〜(h) を，実体・現象・属性・値の区別や係りの方向を無視してすべて適用した場合である．キーワードの意味分類から係り受けの制約を行う本手法に対し，キーワードの意味は考慮せず品詞のみから想定されうる可能なパタンを用いた手法（表5.3の (iv)）と言い換えることができる．ここでは (iv) と公平な対比を行うため，並列構造解析を含まない (ii) と比較を行う．表5.1がセンテンス単位の集計であったのに対し，表5.3はページ単位の集計である．センテンス単位で集計を行った時点で (ii) と (iv) の精度の差は0.5%であったが，5.2.1節で行った網掛け部分についての考察の通り，ページ単位で集計することによって，表5.3において差は1.0%に広がった．

(iii) は表3.2に4.1節「並列構造解析」のアルゴリズムを加えたもので，(ii) と比較するとMAPにおいて有意な差は観察できない[3]ものの，精度・再現率・F値・MAPのすべてにおいて上回っている．

なお，1.2.2節で挙げた関連研究は，基本的にはクエリが文入力で，同じ文型だけを探す方法であり，クエリが (a) の文型なら (b)〜(h) は検出しないため，再現率が低くなることは必然といえる．また，言い換えに対応するために導入しているルールは，本論文で検討した (a)〜(h) の一部にのみ対応したものとなっており，可能な言い換えを網羅したものではないため，やはり再現率の低下を招くと考えられる．

5.2.3 フィルタリングツールとしての性能の検証

フィルタリングツールとしての性能評価を行うにあたり，本来は表構造や見出し構造を合わせて評価すべきと考えるが，今回は本論文で取り扱った範囲内で検証しておくことにする．

ウェブ空間での検索において，ユーザはほとんどの場合，膨大な検索結果の中から上位にランキングされたページしか参照しない [38]．またユーザが検索結果を参照する場合，ランキングの最上位から降順に結果

[2]カイ2乗検定を行い，(i) と (iii) の再現率間に危険率0.1%で有意差があることを確認済みである．

[3]ただし各平均精度について実施したt-検定では，危険率1%で有意であることを確認できている．

を見ていくのが普通である．そのためユーザがより参照し易い，ランキング上位部分に適合文書が並ぶことが望ましい [39]．つまり，検索結果の評価としては，上位にランキングされた文書の検索精度が重要となる．今回の手法は従来の検索手法と合わせて用いることによって，再現率を大きく下げることなく上位の検索精度の向上を図るものである．

そこで，5.1.2 節で述べた，検索結果上位 100 件に対して，我々のシステム（5.2.2 節 (iii) に 3.1.3 節「係り受けパタンの文中における位置」の戦略を加えたもの）でランキングを行う．また同時に近接性に基づいた最小単語距離 [40] によって同じ 100 ページをランキングする．我々のランキングと最小単語距離によるランキングのそれぞれ上位 20 位までの適合ページ数を比較する．

なお最小単語距離とは，2 つのキーワード間に存在する語数を調べ，ページ中の最小値を各ページのスコアとして，スコアの小さいものを上位とする手法で順位を定める方法である．

表 5.4 の左側 4 列は，今回のオープンテストにおける上位 20 件中の適合ページ数の比較である．上位 20 件における適合ページの数が，元の順位 (A) と比較して平均 2 ページ，近接距離 (B) と比較して平均 1.5 ページ，増えた計算になる．また (C) は，表 5.3 において F 値が最も高かった (v) である．これと比較したところ，わずかではあるが本手法の方が適合ページ数が多かった．

一方，右側 4 列は，100 位までの平均精度である．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．最下行は平均精度の平均，すなわち MAP である．表 5.3 及び表 5.4 のそれぞれの手法は，対象データセット及び算出手法が同じであるので，MAP を比較することが可能である．

この表 5.4 の結果より，再現率より精度が重要視されるウェブ検索において，本手法が上位の適合ページ数を最も増やすことができ，フィルタリングツールとして有効であることが示された．

5.2.4 考察

本手法の精度を落とす原因

本手法で不適合ページを拾い上げてしまうケースには，2.2 節で述べた要因のうち，(2-1-1) に該当するものが多く見受けられる．具体的には，見

表 5.4: 提案手法と類似手法の適合ページ数と平均精度

	適合ページ数				平均精度			
	(A)	(B)	(C)	(D)	(A)	(B)	(C)	(D)
(1)	14	18	20	20	0.673	0.888	0.918	0.920
(2)	18	18	20	20	0.859	0.908	0.937	0.937
(3)	8	7	10	11	0.426	0.467	0.525	0.532
(4)	19	13	17	18	0.793	0.719	0.776	0.785
(5)	15	16	16	17	0.684	0.760	0.764	0.792
(6)	8	9	15	15	0.572	0.541	0.592	0.609
(7)	15	18	18	18	0.705	0.858	0.852	0.859
(8)	13	11	14	12	0.591	0.551	0.633	0.594
(9)	16	19	20	20	0.809	0.919	0.913	0.920
(10)	17	17	18	19	0.806	0.841	0.852	0.857
(11)	18	17	19	18	0.809	0.740	0.835	0.821
(12)	20	20	20	20	0.944	0.990	0.988	0.988
(13)	14	13	16	14	0.571	0.657	0.677	0.628
(14)	16	16	16	16	0.838	0.872	0.848	0.847
(15)	7	5	3	6	0.359	0.311	0.241	0.346
(16)	3	3	3	3	0.459	0.586	0.612	0.612
(17)	11	19	17	19	0.672	0.890	0.828	0.880
(18)	12	18	18	18	0.643	0.910	0.868	0.920
(19)	10	12	14	15	0.488	0.542	0.703	0.665
(20)	6	2	2	2	0.546	0.268	0.275	0.276
計	260	271	296	301	0.662	0.711	0.732	0.739

(A)…元のランキング

(B)…最小単語距離によるランキング

(C)…キーセンテンスの有無によるランキング

(D)…本手法のランキング

出しのみで中身（解説）がない「書籍／授業／セミナー紹介」や，個人的な内容が主体の「ブログ」「日記」「体験談」，あるいは過去の内容である「ニュース」「キャンペーン／新製品の告知」など,「検索意図にそぐわないページタイプ」とも言い換えられるものが特に目立つ.

本手法の再現率を落とす原因

ここでは，キーセンテンスが含まれていながら本手法で拾い上げられなかった適合ページについて検討する．これは，本手法の再現率を落とす原因でもあり，かつ，表 5.3 の (v) の精度が予想よりも下がらなかった原因でもある.

そのようなページを個々にチェックしてみたところ，修飾－被修飾関係に無いキーセンテンスは，そのページを適合ページと判断する要因とは無関係であり，そのページを適合たらしめる要因は別に存在するケースが多く見受けられた.

特に多く見受けられたのは，文以外の構造（表構造，見出しの構造など）によって 2 語のキーワードの関係が明示的に示されているケースである（約 4 割）．例えば,「使い方はマスカラの塗り方とまったく同じです.」というセンテンスあったとして，このセンテンス自体は「マスカラの使い方」を説明するものではない（実際の事例では「まつげ美容液の使い方」の説明文であった）．マスカラの使い方の記述は，これとは別の文脈中に表や見出し構造の形（大見出しに「マスカラ」が含まれ，その関連している範囲内の小見出しに「使い方」が出現するなど）で現れ，これが修飾－被修飾関係の役割を果たすということは十分考えられる.

また，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修飾－被修飾関係にあり，その近辺に欲しい情報が存在しており，キーセンテンスはそれと離れた別の文脈の中でたまたま使われていたケースも見受けられた（約 1 割）.

これらのケースに共通して言えることは,「キーセンテンスの存在は，それだけでは適合ページと判断する根拠にはなり得ない」ということである．表構造や見出し構造，あるいは同義語・類義語などで 2 語の関係が表現されている適合ページには，キーセンテンスを含まないものもあり，それらは表 5.3 のいずれの方法でも抽出できていない．従って，本手法の再現率の低下を防ぐために，文の修飾－被修飾関係のチェックを緩めるのではなく，表や見出しの構造，あるいは同義・類義語などで意味的関係が表現されているものを抽出することを考えるべきである．なお，我々は

既に表構造及び見出し構造に関する検討も開始しており，これによって再現率の上昇が見込めることについては，5.3 節および 5.4 節で述べる．

再現率を低下させるもう一つの主要な要因として，3.1.2 節で検討したパタン以外の構文パタンで 2 語の関係が表されている文の存在を挙げることができる（約 4 割）．例えば，年金と解説というキーワードに対し，「年金にまつわる意外な落とし穴を解説」のように，両者の間に複数の自立語が介在する形ではあるが，意味的には「年金についての解説」であることが読み取れる文が使われている例がある．このようなケースの中には，ある程度パタン化して待ち受けることも可能と思われるものもあるが，それによって精度を落とすこともあり，フィルタリングツールとしての本システムの利用法と合わせて今後検討を進めてゆく必要がある．

文中における位置の戦略に対する効果

表 5.3 及び表 5.4 の全ての MAP を比較すると，表 5.4(D) の MAP が最も高く 0.739 であるものの，表 5.3(iii) の MAP は 0.738 であり，戦略の違いである「係りパタンの文中における位置」が有意と言える差はない．

この原因を分析したところ，文中における位置の戦略を適用したときの効果が，高いキーワード対とほとんどないものがあることがわかった．従って，平均してしまうと有意差がでないが，事例によっては効果はあるとみられる．例えば,「女性　政治家」というキーワードの場合，上位 100 件の中に不適合ページにも関わらず係り受けパタンに一致するものが 49 ページあり，うち 24 ページについては文中における位置の戦略によって誤検出を防ぐことができている．

不適合ページの排除効果

2.2 節において，不適合要因の分析結果として「5 割以上の不適合ページが排除できるものと期待できる」と述べた．本手法によって，実際にはどのくらいの不適合ページが排除できたのかを確認する．5.2.2 節の実験に用いたデータ 1995 ページ中，不適合ページは 962 ページ含まれている．表 5.3(iii) では，そのうち 171 ページが修飾－被修飾関係にあると誤判定されているが，残り 791 ページについて修飾－被修飾関係に無いと判断している．すなわち，不適合ページの 8 割以上を排除できた計算となる．

補足実験

実際の検索シーンにおいて，1〜2 語で検索される事例が多いことは事実である．ただし，2 語が選ばれた際に，検索意図が 2 語で概ね伝えられているかどうかは別問題である．もし，現実には適切な 2 語が必ずしも選ばれていないとしたら，5.1.2 節で示したデータセットでは（検索意図を推定しやすいキーワードしか評価対象としていないため），評価が不十分であることも考えられる．そこで，実際に 2 語で検索するケースにおいて，提案手法によるリランキング前後の各上位 20 ページの適合文書数を比較する実験を行う．過去実際に検索した履歴の中から 2 語の事例を選んでもらうところから，検索意図（適合・不適合判定基準）の設定，適合判定の作業までを各被験者に任せることとする．1 人 1 検索課題，計 20 人を対象とする．

実験結果を表 5.5 に示す．Before は Google の上位 20 件，After は Google の上位 100 件に対し提案手法でランキングし直した後の上位 20 件中の適合ページ数を表す．

検索意図（適合・不適合判定基準）とキーワードを照らし合わせると，第 3 者からみて適切とは言いがたいケース（例えば (3) は「Messenger アンインストール」というキーワードであるが「Windows Messenger」のみが適合で「Yahoo！メッセンジャー」や「MSN Messenger」は不適合と判定されているなど）もいくつか存在した．しかしながら，20 個の検索課題の中で，2 個が適合ページを減らしてしまうものの，13 個は適合ページが増加するという結果となった．また，両者の差をウィルコクソンの符号付順位和検定によって検定したところ，危険率 1%で有意差が認められた．

5.3 表構造についての評価実験

本節では，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および評価手法については 5.1 節で述べる．結果の詳細は 5.3.1 節で述べるが，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．

表 5.5: 提案手法によるリランキング前後の適合ページ数

	検索キーワード		Before	After
(1)	伊勢神宮	アクセス	9	11
(2)	オリエンタルラジオ	ブログ	11	14
(3)	Messenger	アンインストール	12	14
(4)	森林公園	浜北	9	5
(5)	浜松市城北	地図	19	19
(6)	ワンピース	着こなし	14	15
(7)	Acrobat	ページ番号	13	14
(8)	Linux	コマンド	18	18
(9)	名古屋駅	ホテル	19	19
(10)	バスケ	ルール	13	16
(11)	DOCOMO	最新機種	10	10
(12)	企業	ランキング	12	13
(13)	日本	異常気象	11	16
(14)	ドラマ	名言	15	15
(15)	浜松	ランチ	13	17
(16)	サッカー	欧州	14	13
(17)	岩手	ツアー	17	18
(18)	履歴書	志望動機	9	14
(19)	小説	新刊	15	18
(20)	プロバイダ	比較	19	20
		計	272	299

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易いキーワードと，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言うと，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 5.3.2 節で述べる．5.3.3 節では，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 3 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．

5.3.1 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，5.1.2 節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，ついで各課題について

表 5.6: 提案手法の精度・再現率・F 値

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 5.7: 適合ページ数と MAP（全データ対象）

	適合ページ数	MAP
(Ⅰ) 元のランキング	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

ランキングを行った結果上位 20 位までの適合ページ数および 100 位までの MAP を算出した.

表 5.6 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 5.7）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.

5.3.2 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，5.1.2 節で述べた 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およびそれらが持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 1 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，表の中のある一箇所に答え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 1 つのセルに答えがあるはず．
(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 5.8）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 5.9 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，元の順位 (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[4]．さらに，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．

[4] (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 5.8: 抽出したキーワード対

	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

表 5.9: 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

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)…元のランキング

(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

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言える．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．

5.3.3 考察

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5.10に示す．なお，このうち要因(2)(3)(5)は，文判定・見出し判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．

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，3.2.2節で述べた「セルの中のキーワードの位置」を考慮する必要がある．

表 5.10: 精度低下の要因

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%

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．

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 5.10 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持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うる．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 3 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 [41].

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持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．

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．

2 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．

5.4 見出し構造についての評価実験

見出し構造については，見出しの抽出精度，支配範囲の精度，検索性能の向上，の 3 つの視点から評価実験を行う．

5.4.1 評価用データの作成

見出し構造の係り受け関係が検索に有効であることを検証するためには，前提として階層関係が正確に解析できていることが重要である．

そこでまず，以下の項目に関して評価実験を行う．

- 見出しそのものが正しく抽出できているか
- 見出しに含まれるキーワード間の関係（先祖－子孫関係）が正しく取れているか

そのためには，人手によって正解が付与された評価用データを用意し，プログラムの実行結果と比較をする必要がある．5.1.2 節で述べたテストコレクションから，以下の要領でページを抽出し,「正解」として見出しと支配範囲を表すタグを埋め込む．

評価用テストコレクションの作成手順

(1) 200 ページの選定を行う．
  (1-1) 5.1.2 節で述べたテストコレクション 118 組から，元の順位の上位 10 ページずつ，計 1180 ページを用意する．
  (1-2) サイズが 0 バイトのページを削除する．
  (1-3) 残りをサイズ順にソートする．
  (1-4) (1-3) の結果により中間の順位の前後合わせて 200 ページを仮候補とする．
  (1-5) (1-4) の仮候補から，ドメインが同じページを削除し，次にサイズが近いページを新たに追加する．
  (1-6) (1-5) の後，キーワード対毎に満遍なくする為，追加・削除を行う（キーワード対毎に多くても 3 ページ）．

表 5.11: 見出しの抽出精度

n=1,954（見出し）

	Hit	FA	Miss	Precision	Recall	F-value
	1578	445	376	78.0%	80.8%	79.4%

(2) 被験者（大学生3名）に，3.3.3節で定義した「見出し」と「支配範囲」を記入してもらう．
  (2-1) 3人が一致した「見出し」や「支配範囲」をテストセット候補とする（200ページ中170ページが一致）．
(3) プログラムにより，上記で記入してもらった「見出し」や「支配範囲」と，本研究の戦略で抽出した「見出し」や「支配範囲」を比較し評価するために，(2-1)のテストセット候補から150ページをランダムに選び別途定義した「見出し」や「支配範囲」のタグを付与する．この中から，クローズドテストセット50ページ，オープンテストセット100ページとし，これらを評価用データとする．

5.4.2 見出しの抽出精度

4.3.1節で述べた手法で，本論文で定義した「見出し」を正しく取り出せるか否かを評価した．その結果を表5.11に示す．表5.11中のHitは正しく抽出できたもの，Missは取りこぼしたもの，FA(Flase Alarm)は誤って抽出したもの，PrecisionはHit/(Hit+FA)，RecallはHit/(Hit+Miss)，F-valueはPrecisionとRecallの調和平均である．

失敗事例を分析したところ，一行全体がリンクや強調のものを「見出し」として取り出すという手法にFAが多く，Missに関してはレイアウト的な特徴やタグでは取れず，意味の理解や繰り返し構造の認識によって抽出が可能となるケースが多いことがわかった．

5.4.3 見出しの2項間関係の精度

4.3.1節で述べた「見出しの階層構造判定アルゴリズム」の精度を評価するにあたり，2つの見出し間の関係（以降，見出し2項間関係）を調査

表 5.12: 見出し 2 項間関係の精度

n=4,023(先祖－子孫関係数)	
Hit	2115
FA（見出しを誤って抽出した分）	866
FA（先祖－子孫関係を誤って抽出した分）	167
Miss（見出しを取りこぼした分）	1358
Miss（先祖－子孫関係を取りこぼした分）	550
Hit+FA(本システムで抽出した先祖－子孫関係)	3148
Hit+Miss(正解の先祖－子孫関係)	4023
Precision	67.2%
Recall	52.6%

することにする．見出し 2 項間関係が先祖－子孫関係になっているか否かについて，精度を検証する．

正解の評価用データとの比較結果を表 5.12 に示す．なおこの比較において，先祖－子孫間の距離は考慮していない．

表 5.12 中の，Hit とはテストコレクションから得られる木からの見出し間の先祖－子孫関係つまり，目視による先祖－子孫関係を本手法においても正しく取り出せているケース，Miss とは目視による先祖－子孫関係を本手法では抽出できていないケース，FA とは目視による先祖－子孫関係では無いものを本手法で先祖－子孫関係として抽出しているケースについて，それぞれ先祖－子孫関係の数で表している．Miss・FA については，それぞれ「見出し」の抽出が原因となっているケースが多いため，分けている．

図 5.1 は「見出し」は正しく取り出せているが Miss となってしまった例である．目視による判断ではページ内の初めに出現する「見出し」①の“ YAHOO!ブックス ”はそれ以降の全ての「見出し」の親となるが，本手法では右の様な木を構成し，③の“ バカの壁/養老孟司/著 ”とは兄弟関係になるため，子ノードとして②の“ トップ ”から“ バカの壁 ”までの 5 つしか持たない．よって，以降の③の“ バカの壁/養老孟司/著 ”や④の“ 書籍画像 ”等とは親子や先祖－子孫にならず Miss となる．①の見出しと③の見出しにおいて，色の違いはあるが，どの色がどの色より上位階層となるということは一様には言えず，階層判断材料に利用できな

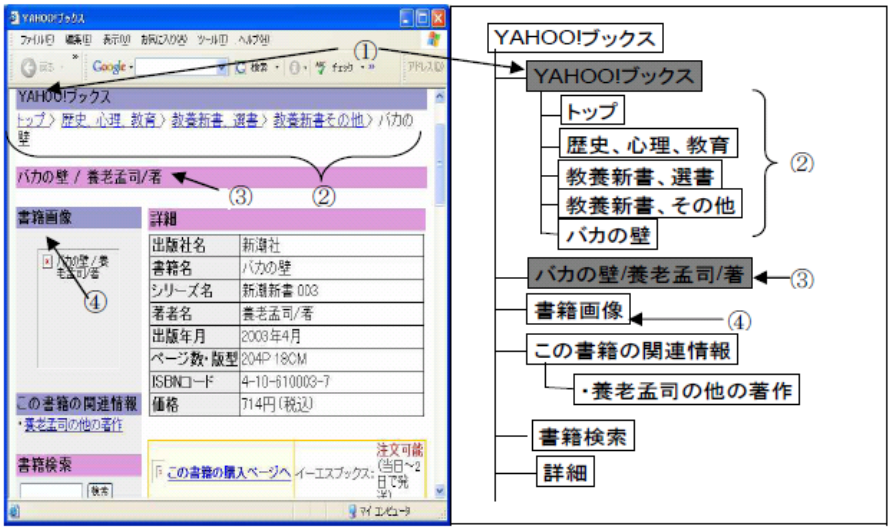

図 5.1: 先祖－子孫関係の抽出失敗例

い．また，見出し内の語を見ることにより，①が③の上位であるとの判断をしていると考えられ，今回の形式のみに着目した手法では正しく取り出せないという問題がある．

5.4.4　検索性能の評価

これまで述べてきた「見出し階層構造判定アルゴリズム」により作成する木構造の先祖-子孫に検索キーワードが 1 語ずつ含まれているか否かで，ページ内の係り受け関係の有無を判断する．これを「見出し判定」とし，5.1.2 節で述べたテストコレクションを用いて検索性能の評価実験を行う．

今回は，文判定と表判定による結果に，見出し判定手法を追加することによる精度の変化を確認する．

表 5.13 は，文判定または表判定で適合と判断されたページを除いて見出し判定を行った結果である．

それに対し，表 5.14 および表 5.15 は，文・表判定に見出し判定を加えた結果である．文＋表判定と比較し，見出し判定を加えた場合においては，F-value などに改善の兆しが見受けられるものの，MAP や上位 20 位の適合ページ数はほとんど変化がなく，満足のいく結果とは言いがたい．

表 5.13: 文・表判定で○のページを除き見出し判定を行った結果

n=6,154(ページ数)

	Precision	Recall	F-value
見出し判定	50.9%	23.4%	32.1%
先祖子孫関係のみ 見出し判定 ＋2つのキーワードが1つの見出し	50.6%	26.5%	34.8%

表 5.14: 文・表判定に見出し判定を加えた結果（精度・再現率・F 値）

n=11,776(ページ数)

	Precision	Recall	F-value
文＋表判定	66.1%	64.7%	65.4%
文＋表＋見出し判定	63.9%	73.0%	68.1%
先祖子孫関係のみ 文＋表＋見出し判定 ＋2つのキーワードが1つの見出し	63.6%	74.0%	68.4%

表 5.15: 文・表判定に見出し判定を加えた結果（MAP・適合ページ数）

n=11,776(ページ数)

	上位 100 件の MAP	上位 20 件中適合ページ数
文＋表判定	0.674	1615
文＋表＋見出し判定	0.680	1615
先祖子孫関係のみ 文＋表＋見出し判定 ＋2つのキーワードが1つの見出し	0.680	1630

5.4.5 考察

見出し判定の精度を向上させるための課題は以下のとおりである．

(1) 一方のキーワードがタイトルに含まれている場合，もう一方のキーワードが見出しや箇条書きに含まれていると，無条件に○と判定してしまう．タイトルの支配範囲を正しく判定する必要がある．

(2) 見出し中の検索キーワードの位置を考慮する必要がある．

(1) については，6.2 節で検討を行う．

(2) については，文構造の場合と同様に，検索キーワードが従属文にしか存在しないケースを排除する方向で，現在調整を行っている．しかしながら，見出し中のすべての検索キーワードが従属文にしか存在しない，というケースは全ページの 2%足らずであり，効果はあまり期待できない．本来，主文にキーワードが存在している場合も，ヘッドからの距離がある場合は，ヘッドに近い語の概念に引っ張られることにより検索意図から外れる可能性があるため，排除の対象とすべきである．しかしながら，ウェブ文書には単語羅列が多く見受けられ，これと複合語を見分けることが難しいため，効果的な戦略を立てるに至っていない．

第6章　精度向上の工夫

同じタイプの係りを持っているとしても，その係り受け関係がページの中でどの程度の重要度であるのかを決める要因は，他にも存在していると考えられる．また，同じキーワードであっても，ユーザが持つ検索意図によって，欲しいページのタイプが異なることがありうる．

6 章では，これらの問題への対応について述べる．

6.1　ページタイプ判定フィルタ

6.1.1　ページタイプ判定の意義

情報検索をする際，ユーザが持つ意図は様々であり，検索キーワードのみでは判断ができないことがある．例えば，ある製品の情報が欲しいと考え，製品名で検索するとする．そのとき，性能や販売価格などが知りたい場合もあれば，実際にその製品を利用した人の感想などが知りたいこともありうる．すなわち，同じキーワードであっても，意図によって欲しい情報は異なるため，一方を無条件に排除する戦略は好ましくない．ただし，客観的な事実，もしくは主観的な意見のどちらかのみを必要としているユーザーに対して，不要なページを排除することを選択できるようにすることには意義がある．同様に，情報の価値が時間の経過によって変動する（古くなると役に立たなくなる）内容のページ（ニュース，オークション，求人情報 etc.）についても必要／不必要を状況に応じて選べるとよいと思われる．

そこで，各ページに対し「主観的な意見」「時間の経過で価値が減退する情報」が書かれている量をそれぞれ推定し，検索結果の画面に「主観／客観性」「価値減退可能性」としてパーセンテージで表す「ページタイプ判定フィルタ」を検討する．ページタイプ判定フィルタは，検索時にユーザが意図に沿って選択し，利用できる形で提供したいと考えている．

これにより，検索結果画面からリンクをたどらないとわからなかった「主観的な意見」か「時間の経過で価値が減退する情報」かといったおまかな判断が検索結果画面上のみでできるようになり，求めているデータにたどりつくまでの時間や手間を短縮する事ができる．

ウェブページをカゴテリごとに分類する手法はこれまでにも多数提案されている．中でも我々と最も近い研究に,「事実」と「意見」を分類する Finn らによる手法 [42] がある．Finn らは，bag of words や品詞，記号の数や文の平均長といった統計情報，さらに人手で作成した主観性を表す語のリストなどを用い，機械学習によって分類を実現している．この手法では「フットボール」や「政治」といったドメインを限定した中では高い精度で判別できるものの，学習させたアルゴリズムを異なるドメインに適用すると，精度が著しく落ちるという問題が示されている．一方我々は助動詞などの文末表現を利用することにより，ドメインに依存しない判定を実現する．文末の助動詞から時制（テンス），相（アスペクト），法（モダリティ），態（ヴォイス），さらに打消しや，疑問形の文かどうかといったことを分析し，ページタイプの特徴をつかむことを試みる．また，文末の「思考」や「表明」を意味する動詞（以降「思考・表明動詞」と呼ぶ）と終助詞にも着目し，助動詞同様，タイプによる分析を行う．

体験記など（主観的な表現が多いページ）は「てみた」「てきた」という表現が多いであろうし，一方で公式サイトなど（客観的な表現が多いページ）には「ございます」「おります」などの敬語が多いと思われる．また，ニュース記事や日記（古くなると価値が無くなるページ）は過去のことについて書かれることが圧倒的に多いので，必然的に過去の助動詞「た」が多く出現すると推定できる．

6.1.2　文末に着目する理由と分類手法

日本語は思考，状態，意志，態度など，文末に重要な表現が多くあらわれる [43]．たとえば『先日発売されたメガマックは好評である』という文をみると「た」は過去の助動詞だが，文全体としては過去の話ではない．このように文末でないところの表現は本筋ではない場合が多い．そこで，文の終端からさかのぼって自立語が出現するまでを文末とし，その文末中から終助詞，助動詞，思考・表明動詞を抽出する．抽出した語ごとに頻度を算出し，あらかじめ重回帰分析により得た式に当てはめ「主

観／客観性」「価値減退可能性」を計算する．

6.1.3 評価実験

教師データの作成

まず，2007年7月某日のYahoo!検索ランキング[1]から以下の8種類の検索ワードを選んだ．

- 参議院議員選挙
- google
- 執事喫茶
- 牛丼復活
- mixi
- ブログ
- ビリーズブートキャンプ
- 土用の丑の日

これらの各検索ワードに対して100ページ分の検索結果を取得する．次に，取得したそれぞれのページを被験者に閲覧させ，その内容に対して「客観的である」を1,「やや客観的である」を2,「やや主観的である」を3,「主観的である」を4とした4段階評価を行なってもらう．1ページにつき2人が評価を行い，平均をとったものを正解データとする．意見が2段階以上分かれた場合は，判断を第3者に委ねる．なお，800ページ中，ページの移動などにより取得できなかった42ページを除いた，758ページを実験データ（表6.1）とする．

重回帰分析

作成したプログラムを用いてウェブページ中の終助詞，助動詞，思考・表明動詞の頻度をとり，目視での評価と特定の語の出現頻度に相関関係があるか否かを調査した．今回は，実験データ758ページから半数をランダムに抽出し，教師データとして重回帰分析を行った．終助詞，助動詞，思考・表明動詞は，400ページ中に40回以上出現しているものを対象とした．次に目視で評価した正解データを従属変数，400ページ中の終助詞，助動詞，思考・表明動詞の種類ごとの頻度を独立変数とし，ステッ

[1] http://searchranking.yahoo.co.jp/

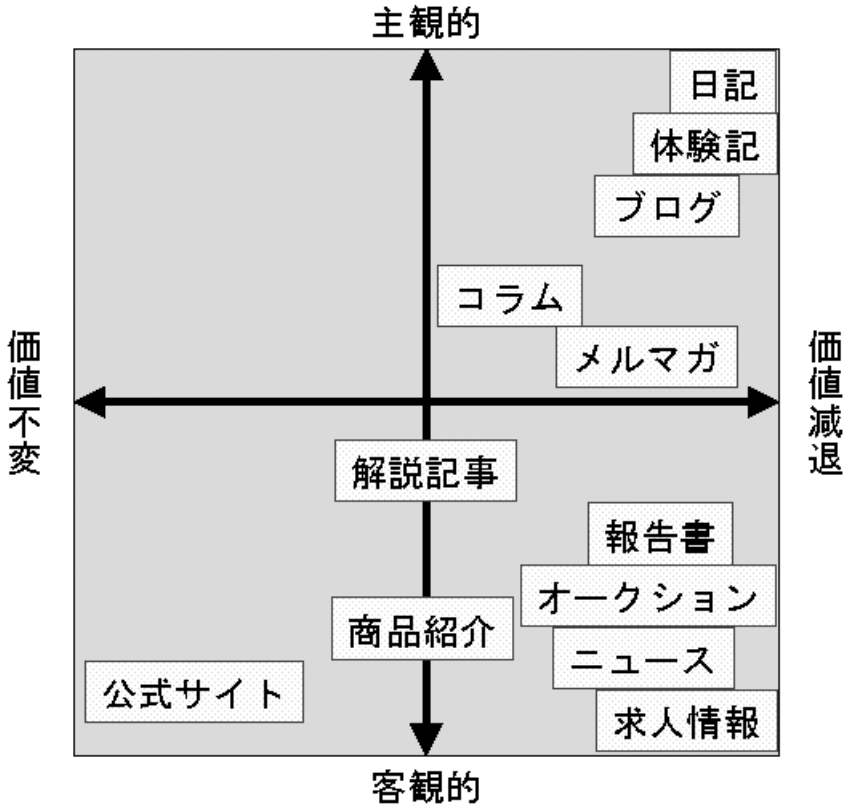

図 6.1: 「価値減退可能性」と「主観性」

表 6.1: 目視データ内訳

主観／客観度	ページ数	割合	
1	340	44.9%	客観
1.5	64	8.4%	↑
2	64	8.4%	
2.5	36	4.7%	
3	41	5.4%	
3.5	66	8.7%	↓
4	147	19.4%	主観
計	758	100.0%	

表 6.2: 偏回帰係数

	偏回帰係数		標準誤差	t 値	p 値
発表する	-0.233	客観	0.083	2.817	0.005
た	0.031	主観	0.008	3.874	0.000
できる	-0.243	客観	0.057	4.286	0.000
コメントする	0.235	主観	0.041	5.728	0.000
分かる	-0.353	客観	0.064	5.539	0.000
ない	0.049	主観	0.019	2.582	0.010
ことができる	-0.148	客観	0.070	2.109	0.036
定数項	1.967		0.068	28.785	0.000

プワイズ法を用いて変数を絞り込み，重回帰分析を行った．偏回帰係数を表 6.2 に示す．従属変数の値は，大きい方が主観的であることを意味する．分散分析を用いて重回帰式の検定を行った結果，F 値は 17.2 であり，有意水準 5%において帰無仮説が棄却されるので，ウェブページの客観，主観表現の判別に文末表現を利用することが有効であるという結果を得ることができた．

この実験で得られた回帰式を，残りの 400 ページに当てはめたときに得られる結果を，今後は詳細に分析していく．

また，意味の近い語をグループ化したり，過去の助動詞「た」と組み合わせた場合の出現傾向を見るなどして，さらに判定に効果的な要因を探していきたい．

6.1.4 分析結果

分析結果として主なものを以下に挙げる．

- 助動詞
  - ○ センテンスごとに集計
    - 文末に「た」「ことにする」「らしい」という助動詞が存在する文は，価値減退可能性の高いページである確率がいずれも 90%以上
    - 「ことにする」は主観性も約 90%と高く，日記である確率も約 60%

- 「ことにする」＋「た」(「ことにする」より後ろに「た」が出現する）は価値減退可能性，主観性ともに 95%以上とひときわ高く，また日記と体験記がそのほとんどを占める

○ ページごとに集計

- 文末に「た」の文が 1 つでも存在しているからといって価値減退可能性が高いとは限らない（約 65%）が，日記は 22 ページ全てに「た」が存在
- 「てくる」は日記 22 ページ中 20 ページに存在
- 「ことにする」はページ中に 1 つでも存在しているページは価値減退可能性が高い可能性は 90%，主観性が高い可能性は 80%，また「ことにする」が存在する 38 ページ中，体験記が 17 ページと半分近くを占める
- 「らしい」も価値減退可能性が高いページである確率が約 90%でとても高い
- 日記のほとんどには「た」だけでなく「ました」(20/22 ページ)「ていた」(20/22 ページ）も含まれている
- 「ことにする」＋「た」はページ中 1 つでも「〜ことにした」と出てきたら 90%以上価値減退可能性が高いページ
- 「ことにする」＋「た」は体験記である確率が約 50%，日記が約 30%，センテンス集計のときと割合は逆転しているもののほとんどを占めることに変わりはない，逆転したのは日記は 1 ページに何日分も書かれることが多いため

○ 割合（1 ページ中の該当の助動詞が含まれる文の数／全ての文の数）で集計

- 1 ページ中に含まれる「た」の平均出現数は 8.7 文で全体の 4.6%（ページごとの割合の平均），それが価値減退可能性 1 のページであると 0.6%になり，価値減退可能性が 4 のページであると 10.3%，主観性が 1 のページであると 2.7%に対し主観性が 4 のページであると 12.6%と 1 ページ中に含まれる量に歴然と差がある
- 他に割と顕著な差があるものは「ている」「てくる」「てしまう」など
- 先ほど価値減退可能性，主観性ともに高かった「ことにする」はそれほど差が見られる訳ではなく，1 つのページ中に数多く出現する助動詞ではない性質であることがうかがえる

- 「ます」＋「た」は 1 ページあたりの出現数は平均 4.1 文で全体の 1.9%，それに対し価値減退可能性 1 の場合，出現確率は 0.4%，価値減退可能性 4 のときは 4.0%，主観性 1 なら 1.0%，主観性 4 なら 5.6%である

○ その他

- 願望の「たい」は広告のキャッチコピーなどによく利用されているので価値減退可能性・主観性は低い
- 「てみる」「よう」は勧誘の意味が広告，紹介，リンク集などに含まれ易くなっている
- 「くださる」「いただく」といった敬語は価値減退可能性，主観性の高いページにはあまり出現しない傾向にある
- 過去形に限定して集計をした結果と限定していないものとを比較すると過去形に限定することで最も特徴が現れたのは「ます」

・ 思考・表明動詞

○ センテンスごとに集計

- 価値減退可能性と関連がありそうなものは「思う」(86.4%)「感ずる」(85.1%)
- 過去形の文に限れば「思う」「取る」「発表する」＋「た」

○ ページごとに集計

- 価値減退可能性と関連がありそうなものは「発表する」(81.3%)「期待する」(81.0%)
- しかも「発表する」はニュース記事である確率が 52.5%
- 過去形の文に限れば「思う」「発表する」＋「た」

○ 割合で集計

- 価値減退可能性と関連がありそうなものは「思う」「取る」「発表する」「感ずる」「期待する」
- 主観性と関連がありそうなものは「思う」「取る」
- 日記と関連がありそうなものは「取る」
- 記事と関連がありそうなものは「発表する」「期待する」
- 体験記と関連がありそうなものは「取る」「目する」「感ずる」
- 過去形の文に限れば「思う」「見る」「言う」「発表する」＋「た」

このように，ページタイプによって文末表現は様々な特徴を持って表れる傾向にあり，文末表現からページタイプを推定することはある程度期待できると思われる．

6.2　ページ内の主要部の特定

ページ内において，検索キーワードが上部に出現する場合と，末端部にて付帯的な記述の中にのみ存在しているような場合とでは，ページにおける語の重要度が異なることは明らかである．そこで，ページ全体中のキーセンテンスの位置を求め，誤った適合ページの検出を抑える手段として利用する．

6.2.1　ページ内の主要部の特定の意義

ウェブページの特徴として，１つのページ内に多様な情報が含まれていることが挙げられる.「本文」と呼べる主要な情報から，サイトのメニューや関連するリンク集，広告などまでが１ページ内に納まっていることが少なくない．これでは，検索キーワード間の意味的関係が強くても，検索意図にそぐわないページがあっても不思議ではない．本研究の表判定や見出し判定では，ページタイトルの支配範囲は，ページ全体であると見なされる．しかし，現実には，一方の語はページタイトルにあるものの，もう一方の語は本文とはなんら関係のない広告の中に存在し，係り受け関係にあるとは言い難い出現の仕方をすることがある．そこで,「ページタイトルの支配範囲」を正確に特定する手法を検討する．

6.2.2　ページ内の主要部の特定の手法

ページ主要部の特定を行う前に，まずはページの分割を考える必要がある．服部ら [44] は，タグの階層が大きく変換する部分をブロックの区切りとみなす手法で，これに成功している．そこで，これに倣って分割地点を求める．

次に，分割後の各ブロックの中で，メインの情報が書かれているものを探す．そこで，ページタイトルが，メインの情報の内容を端的に表わしている可能性が高いことに着目し，ページタイトルを構成する自立語が，各ブロックごとに何種類含まれているかを確認する．最も多くの種類が含まれているブロックを，タイトルの表す内容に近い，すなわち主要部であるとみなす．

なお，現在，ブロック区切りの判断材料として，文末表現も利用することを検討している．広告には勧誘表現が多いであろうし，メニューは短い名詞句の形をしていることが予想されるからである．

主要部を特定したら，そこに係り受け関係にある検索キーワードが存在するかどうかをチェックする．主要部に存在しなかった場合は，ページ全体における重要度が低いと見なせる．あるいは，ページタイトルの支配範囲外であって，実はページタイトルと係り受け関係には無い，という可能性もある．そこで，今後，ランキングのスコアに差をつけることを検討し，検索性能の向上につなげていきたい．

以下に，ページ内の主要部を特定するためのアルゴリズムを詳述する．

(1) ページタイトルを構成する語を抽出する．
    1. ページタイトルを形態素ごとに分ける．
    2. 自立語に該当する形態素を抽出する．

(2) ウェブページの分割を行う．
    HTML ファイルを読み込み，ファイルの先頭から以下のアルゴリズムに従って処理する．
    1. 始まりタグならば，PUSH する．Ex. <HTML>, <TABLE>, <DIV>, <IMG>
    2. 終わりタグならば，POP する．Ex. </HTML>, </TABLE>, </DIV>
    3. スタックに積まれたタグのうち，4 分の 1 以上が連続して POP されたら，次に PUSH された時にページを区切る．
    4. <HR>タグ（水平線）ならば，その場所でページを区切る．

(3) HTML 文書からテキスト（及び画像の alt オプション）を抽出する．
    1. テキスト（および画像の alt オプション）を区切られた範囲ごとに抽出する．

(4) 主要部分を特定する．
    1. (3) のテキストにいくつ（何種類）タイトルワードが含まれているか調べる．
    2. 異なり数が最も多い箇所を主要部分と定める．

6.2.3 仮説の検証：予備実験の結果および考察

今回，予備実験として，検索キーワードを「検索エンジン　比較」とし，その検索結果 100 ページにおいて，ページタイトルを構成する語の分布状況を確認した．

その結果，主要部分の特定精度（図 6.2）は $(69+10)/(100-11) = 88.8\%$ となり，本戦略が有効である手ごたえを得られた．

うまくいかなかった事例は，その多くが下記のようなケースに該当した．

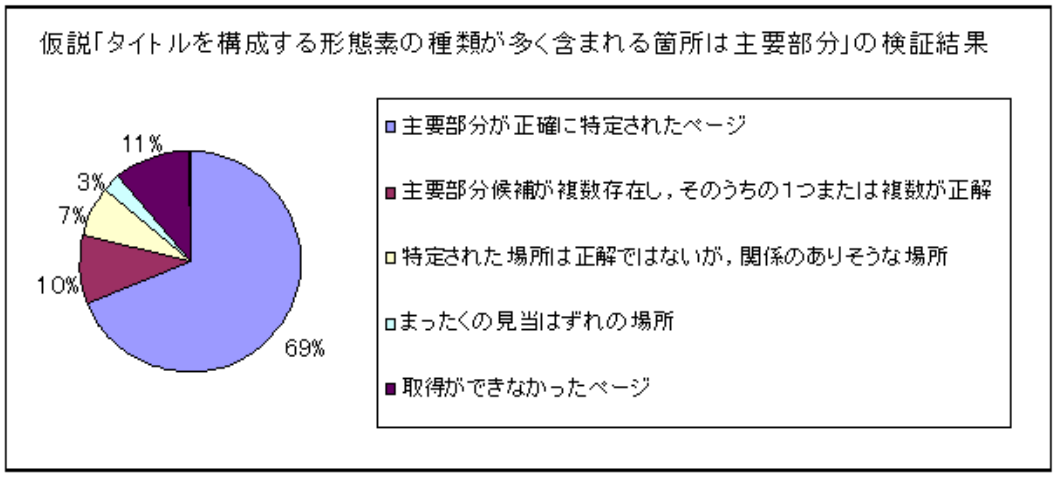

図 6.2: タイトルを構成する形態素による主要部分の特定精度

- 図 6.3 のように，メインではない場所に，ページタイトルがそのまま本文中に出現する場合がある．
- ページタイトルがわずかな語で構成されているケース．

対応策としては，形態素の種類数のみならず，語の頻度や文字の大きさ，情報の量なども判断材料に含めることが考えられる．

今後は大規模に実験を行い，効果を確認する．

図 6.3: 主要部分特定の失敗事例

第7章　結論

本研究では，2語キーワードによる検索において，文や表，見出し構造の係り受け関係に着目することで検索精度を向上させることに効果があることを示した．7章では，その成果と今後の課題について述べる．

7.1　成果

本論文では，従来の自然言語処理の背景にある語の意味や依存関係等に関する考え方を整理し，それに基づいて情報検索手法に関する1つの提案を行った．

ウェブサーチエンジンを利用する際，検索キーワードとして複数の語を入力する場合において，それらのキーワードがウェブページ中に修飾－被修飾関係をもって出現すれば，検索意図に適ったページである可能性が高い．この仮説に基づき，検索性能の向上を目指した．ウェブの多様な文書構造に対応するため，文構造・表構造・見出し構造から修飾－被修飾関係を抽出する手法について，それぞれ検討した．

さらに，精度を向上させるための工夫として，主観／客観的なページ，時間の経過によって価値が減退するページの特定や，ページ内の主要部分の特定を試みた．

提案手法の有効性を評価するために，既存の検索エンジンのフィルタリングツールとしてシステムを構築し，評価用データセットを作成した．これを用いて評価実験を行ったところ，提案システムが元にした検索エンジンと，文判定・表判定・見出し判定を合わせた場合とを比較すると，118クエリの上位20件の適合ページ数が平均1.71ページ増加することが確認された．

これによって抽出した修飾－被修飾関係の妥当性が示された．

7.2 今後の課題

7.2.1 現状の改善

- **各構造の抽出精度の改善**
  文・表・見出しの各構造とも，HTML ファイルから正確に構造を抽出する必要があるが，HTML の記述の自由度は高く，ウェブページの製作者の好みにより様々な特徴を持つあらゆるウェブページに対応することは困難である．しかしながら，本手法の有効性を論じるうえで，一定の水準の精度は必要不可欠である．今後，
  - 文構造については，文の境界を正しく抽出すること，
  - 表構造については，表／レイアウトの判断，見出しの判断の精度向上と，タイプ (i) とタイプ (ii) の判断を加えること，
  - 見出し構造については，見出しとその支配範囲の抽出精度を向上させること，

  について，取り組む予定である．
- **スコアリング戦略の改善**
  現在のスコア算出アルゴリズムは暫定的なものであり，キーワード対のタイプを考慮することもなく，文と表，見出しの係り受けの有無を同じ重みで評価しているなど，不十分なものとなっている．また，6 章で触れた精度向上の工夫もスコアリングに反映させるなど，まだまだ改善できる余地は多い．今後さらに詳細な分析を行い，それに基づいた戦略を立てることでより効果的な順位の並べ替えを行うことができると考えられる．

7.2.2 新たなる挑戦

2007 年現在，日本のインターネット利用者は 8 千数百万人と言われ，今もなお利用者人口は増え続けている．利用目的は様々であるが，目的の情報にたどり着くためにサーチエンジンの利用は必須であり，需要が無くなることはない．その一方で近年，ウェブ上では，国内・海外を問わず「Web2.0」関連と呼ばれるサービスが続々と登場している．このような新しい技術やコンテンツが生まれるたびに，サーチエンジンは対応と進化が求められる．今後もウェブサーチエンジンを研究対象としていく価値は充分にあると考える．

これまで，サーチエンジンの検索性能を向上させる手法について研究を行ってきた．今後は検索性能向上のみならず，特定の用途に特化したユニークなサーチエンジンや,「かゆいところに手が届く」知的なサーチエンジンの研究・開発を目指していきたい.

謝辞

本研究を行うにあたり，丁寧な御指導と適切な御指示，御助言を頂きました，静岡大学の伊東幸宏教授，小西達裕准教授，言語情報処理研究所の高木朗殿，国立情報学研究所の小山照夫教授，中京大学の三宅芳雄教授に深く感謝致します．また，同じ研究チームで苦楽を共にした西口直樹氏，伊藤慎一氏，池田彰吾氏，山本晋太郎氏を始め，伊東・小西研究室の学生諸氏には多大な御協力を頂きました．

また，仕事と学業を両立するにあたり，サレジオ高専の市村洋教授ならびに東京高専の教職員の皆様には多大なご支援を賜りました．

以上の方に論文の末尾ではありますが，改めて感謝の意を表します．

参考文献

[1] 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. In *Pre-meeting Lecture at the NTCIR-3 Workshop*, 2002.

[2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *the Journal of the ACM*, 1999.

[3] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proc. of 7th World Wide Web Conference*, 1998.

[4] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. 1998.

[5] Jeffrey Dean and Monika R. Henzinger. Finding Related Pages in the World Wide Web. *Proc. of 8th World Wide Web Conference*, 1999.

[6] 大久保雅且, 杉崎正之, 井上孝史, 田中一男. WWW 検索ログに基づく情報ニーズの抽出. 情報処理学会論文誌, Vol. 39, No. 7, 1998.

[7] Tomek Strzalkowski and Karen S. Jones. NLP Track at TREC-5. In *The Fifth Text REtrieval Confernce(TREC-5)*, pp. 97–102. NIST Special Publication, 1996.

[8] 新美和彦, 兵藤安昭, 池田尚志. 係り受け関係を用いる高精度全文検索. 情報処理学会第 55 回全国大会講演論文集, Vol. 2, pp. 350–351, 1997.

[9] 新美和彦, 兵藤安昭, 池田尚志. 係り受け情報を用いた全文検索とその評価. 「ディジタル図書館」ワークショップ 第 11 回, 1998.

[10] Tsunenori Mine, Hiroki Fujitani, and Makoto Amamiya. A Japanese Information Retrieval Method Using Syntactic and Statistical Information. In *NLPRS2001*, pp. 429–434, 2001.

[11] 藤谷洋樹, 峯恒憲, 雨宮真人. 係り受け情報や語の意味情報，出現確率情報を利用した情報検索手法の提案と評価. 情報処理学会九州支部 火の国情報シンポジウム 2001, pp. 39–46, 2001.

[12] 清田陽司, 黒橋禎夫, 木戸冬子. 大規模テキスト知識ベースに基づく自動質問応答 ーダイアログナビー. 自然言語処理, Vol. 10, No. 4, pp. 145–175, 2003.

[13] 清田陽司, 黒橋禎夫, 木戸冬子. 自動抽出した換喩表現を用いた係り受け関係のずれの解消. 自然言語処理, Vol. 11, No. 4, pp. 127–145, 2004.

[14] Bernard Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, Vol. 36, pp. 207–227, 2000.

[15] 風間一洋, 原田昌紀. Web サーチエンジン技術の高度化. 人工知能学会誌, Vol. 16, No. 4, pp. 503–508, 2001.

[16] Akira Takagi, Hideki Asoh, Yukihiro Itoh, Makoto Kondo, and Ichiro Kobayashi. Semantic Representation for Understanding Meaning Based on Correspondence Between Meanings. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 10, No. 6, pp. 876–912, 2006.

[17] 亀井孝, 河野六郎, 千野栄一. 言語学大辞典, 6 術語編. 三省堂, 1996.

[18] 斎藤秀三郎, 松田福松. 名詞用法詳解. 吾妻書房, 1956.

[19] R. Zanibbi, D. Blostein, and J.R. Cordy. A Survey of Table Recognition: Models, Observations, Transformations and Inferences. *International Journal on Document Analysis and Recognition 7*, No. 1, pp. 1–16, 2004.

[20] 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, Vol. 2003, No. 23, pp. 137–144, 2003.

[21] 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, Vol. 2006, No. 1, pp. 43–48, 2006.

[22] 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.

[23] 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.

[24] 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, Vol. 2003, No. 4, pp. 65–72, 2003.

[25] 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.

[26] 新里圭司, 鳥澤健太郎. HTML 文書からの単語間の上位下位関係の自動獲得. 信学技報, Vol. 103, No. 408, pp. 27–34, 2003.

[27] 松本吉司, 高橋哲朗, 乾健太郎, 松本裕治. Web ページのテキストセグメント階層構造の抽出. 言語処理学会年次大会発表論文集, Vol. 11, pp. 49–52, 2005.

[28] 鈴木泰裕, 高村大也, 奥村学. Semi-Supervised な学習手法による評価表現分類. 言語処理学会 第 11 回年次大会, pp. 668–671, 2005.

[29] 鈴木泰裕, 高村大也, 奥村学. Weblog を対象とした評価表現抽出. 人工知能学会 セマンティックウェブとオントロジー研究会 SIG-SWO-A401-02, 2004.

[30] 藤村滋, 豊田正史, 喜連川優. 文の構造を考慮した評判抽出手法. 電子情報通信学会 第 16 回データ工学ワークショップ DEWS2005, 2005.

[31] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[32] Y. Wang and J. Hu. Detecting tables in HTML documents. In *LNCS*, Vol. 2423, pp. 249–260. Springer-Verlag, 2002.

[33] 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. 情報処理学会論文誌, Vol. 47, No. 5, pp. 1530–1537, 2006.

[34] 大西香織, 田島敬史. Web上の表データの論理構造の発見. *Proceedings of Data Engineering Workshop*, 2006.

[35] 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ー被修飾関係の詳細な分析に基づくWWW検索性能の向上. 情報処理学会論文誌, Vol. 48, No. 10, pp. 3386–3404, 2007.

[36] 久野高志, 安形輝, 石田栄美, 上田修一. Web ページのタイプ判定法. 2000年度日本図書館情報学会春季研究大会発表要綱, pp. 55–58, 2000.

[37] Akiyo Matsumoto, Tatsuhiro Konish, Akira Takagi, Teruo Koyama, Yoshio Miyake, and Yukihiro Itoh. A Filtering Tool for WWW Search Engines based on Semantic Relation between Input Keywords. *Pre-proceedings of 14th European - Japanese Conference on Information Modelling and Knowledge Bases*, Vol. I, pp. 75–88, 2004.

[38] Amanda Spink, Bernard J. Jansen, Dietmar Wolfram, and Tefko Saracevic. From E-Sex to E-Commerce: Web Search Changes. *IEEE Computer*, Vol. 35, No. 3, pp. 107–109, 2002.

[39] 大塚崇志, 山名早人. Webサーチエンジンの新しい評価手法. 電子情報通信学会 第14回データ工学ワークショップ DEWS2003, 2003.

[40] 田馳, 手塚太郎, 小山聡, 田島敬史, 田中克己. 質問キーワードの近接性と密度分布に基づくウェブ検索の改善手法. 日本データベース学会 Letters, Vol. 5, No. 1, pp. 113–116, 2006.

[41] 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.

[42] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Genre Classification and Domain Transfer for Information Filtering. *Proceedings of*

ECIR-02, 24th European Colloquium on Information Retrieval Research, 2002.

[43] 東照二. 歴代首相の言語力を診断する. 研究社, 2006.

[44] 服部元, 松本一則, 菅谷史昭. タグの深さを利用したコンテンツ間距離に基づく Web ページの自動分割方式. 日本データベース学会 Letters, Vol. 4, No. 1, pp. 149–152, 2005.

関連発表

A. 論文

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，情報処理学会論文誌，Vol.48，No.10，pp.3386-3404，2007.
2. 松本 章代，小西 達裕，高木 朗，小山 照夫，三宅 芳雄，伊東 幸宏：表構造における意味的関係に基づく WWW 検索性能の向上，電子情報通信学会論文誌 D，Vol.J91-D，No.3，2008.（採録決定）

B. その他の論文

1. Akiyo Matsumoto, Tatsuhiro Konish, Akira Takagi, Teruo Koyama, Yoshio Miyake, Yukihiro Itoh: A Filtering Tool for WWW Search Engines based on Semantic Relation between Input Keywords，Information Modelling and Knowledge Bases XVI，Edited by Yasushi Kiyoki, Benkt Wangler, Hannu Jaakkola, Hannu Kangassaol, pp.75-88，2005.

C. 口頭発表など

1. 伊藤慎一，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：適応型サイトマップの作成とそのオンラインショッピング支援への応用，情処研報 2007-DD-60，Vol.2007，No.34，pp.57-82 (Mar. 2007).
2. 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 (Mar. 2007).

3. 松本章代，西口直樹，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：表構造における意味的関係に基づく WWW 検索精度の向上，情処研報 2006-DD-55，Vol.2006，No.58，pp.5-11 (May 2006).
4. 松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：文構造における検索キーワード間の修飾－被修飾関係に基づく WWW 検索精度の向上，信学技報，Vol.105，No.595，NLC2005-115，pp.7-12 (Feb. 2006).
5. 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，Vol.105，No.595，NLC2005-114，pp.1-6 (Feb. 2006).
6. Akiyo Matsumoto, Tatsuhiro Konish, Akira Takagi, Teruo Koyama, Yoshio Miyake, Yukihiro Itoh：A Filtering Tool for WWW Search Engines based on Semantic Relation between Input Keywords，Pre-proceedings of 14th European - Japanese Conference on Information Modelling and Knowledge Bases，Volume I，pp.75-88，held in Sweden (Jun. 2004).
7. 松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：ページ内の意味的係り関係に基づく WWW ページ検索結果の絞り込みについて，信学技報，Vol.103，No.408，NLC2003-38，pp.19-25 (Nov. 2003).