Vol. 48 No. 10 情報処理学会論文誌 Oct. 2007

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

松本章代[†1,†2] 小西達裕[†1] 高木朗[†3]
小山照夫[†4] 三宅芳雄[†5] 伊東幸宏[†1]

ウェブ検索エンジンに，ユーザが検索キーワードとして2つの語を入力した場合に，その2語の修飾–被修飾関係を意味と文法カテゴリの両面から詳細に分析し，特定の関係が文書中に出現しているか否かを判定することにより，ウェブ検索エンジンの性能を向上させる手法を提案する．どのような関係を使うことが有効になるかを判定する基礎として，どのようなキーワードが実際に用いられるのかの検索の実態の調査を行い，その結果を用いて有効な関係を選ぶなどを手法の実現に反映させた．提案手法をフィルタリングツールとして構築し，評価実験を行った結果，単なる修飾–被修飾関係を用いる検索手法に比べ，精度，再現率ともに向上した．また，広く使われている検索エンジンを使った上位20位における適合ページ数の実験でも，適合ページ数が平均1.5～2ページほど増えることが示され，この手法の有効性が確かめられた．

Improvement in Performance of WWW Search Engines Based on Detailed Analysis of Dependency Relation between Input Keywords

Akiyo Matsumoto,[†1,†2] Tatsuhiro Konishi,[†1] Akira Takagi,[†3]
Teruo Koyama,[†4] Yoshio Miyake[†5] and Yukihiro Itoh[†1]

In this paper, we propose a method to improve performance of WWW search engines. We focus on the case that two keywords are input to web search engine. We think that the keywords input by users have a semantic relation. We try to improve retrieval accuracy by checking whether two keywords have a dependency relation in the candidate pages or not. We construct a filtering tool which accept output of an ordinary search engine and select plausible ones by checking dependency relation. We show an experimental evaluation of our method. As a result, we compared our method with ones using just direct dependency relation and showed to raise the precision and recall. Furthermore, the best 20 pages by our system contained about 1.5–2 more relevant pages than general search engines. Therefore, we could confirm the validity of our method.

1. はじめに

インターネットの普及にともないキーワード検索によるウェブ検索エンジンは日常的に広く用いられて，現代社会の中でなくてはならない存在になっている．その性能が少しでも向上することの意義は，多くの人が使用していることもあり小さくない．検索されたページの中で適合性の高い順に表示するランキングの性能向上についての基本的なアプローチは，語の出現に関するTF-IDFなどの統計量や語間の距離などに基づくものが主流である．そのような方法では，用いられるキーワードがどのような意味分類に属し，どのような構文構造を形作るかという文法的関係などの詳細な情報を用いていない．このような情報を利用すれば，ユーザの検索意図に合っていると推定できるページを優先的に表示させることが可能となり，検索エンジンの性能を向上させられる可能性が高いと考えられる．検索クエリを自然語文で入力させるシステムには，入力文の係り受け構造，すなわち語間の文法的関係を利

†1 静岡大学
Shizuoka University
†2 東京工業高等専門学校
Tokyo National College of Technology
†3 言語情報処理研究所
NLP Research Laboratory
†4 国立情報学研究所
National Institute of Informatics
†5 中京大学
Chukyo University

用しているものも見受けられる[1)~6)]．しかしながら，検索クエリを文で与えて文書から同形の係り受け構造を探すと再現率の低下は免れない．我々は，検索クエリとしていくつかのキーワードを入力する方式を想定し，キーワードの意味分類を利用して可能な係り受け構造を推定するという方法を試みる．

本論文では，複数の検索キーワードがそれぞれどのような意味分類に属するときに，どのような文法的関係を想定すべきかを詳細に分析し，それらの関係を利用することによって，入力されたキーワードから推定されるユーザの検索意図に適合すると判定されるページを優先的に提示する手法を実現し，その効果を確かめたので報告する．

本研究では扱う対象を検索キーワードが 2 語の場合の修飾–被修飾関係に限定しているが，次の 2 章ではこのことの妥当性を理論的な背景と実際にどのような検索キーワードが使われるのかについて調べた実験の結果を基に論じ，また関連研究について述べる．3 章では，2 語の検索キーワードの各々がどのような意味分類に属する場合に，どのような係り受け構造を想定すれば検索性能の向上につながるかについて検討する．4 章では，システムを実装するにあたり不可欠な，ウェブ文書の構文解析精度を向上させるための対処法について示し，5 章で実装したフィルタリングツールを紹介する．6 章で，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について述べる．

2. 2 語の修飾–被修飾関係を利用した検索

2.1 本手法の適用範囲

2.1.1 2 語と修飾–被修飾関係を使うことの妥当性

本論文では，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．我々は，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．したがって，それらの語が文書中に存在するということだけでなく，それらの語の間の意味的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．

本論文では，語間の意味的関係を表現しうる構造として，係り受け構造をとりあげて考察する．一般には，語と語の意味的関係を規定する構造には，係り受け構造以外にも，表構造，見出し構造，あるいは単語が埋め込まれている図の構造など，多岐にわたる．本論文では，そのうち，最も出現する比率が高いと考えられる係り受け構造に限定して検討を行う☆．

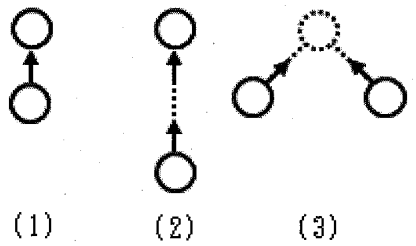


図 1 2 語の構造
Fig. 1 Dependency structures of a phrase or clause containing two keywords.

また，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合せとしてとらえることができること，Jansen ら[7)] や風間ら[8)] の報告のように実際のウェブ検索エンジンにおいて 1 語ないし 2 語で検索されるケースが圧倒的に多いこと，による．

係り受け構造で関係付けられている 2 語の意味的関係の 1 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．たとえば，「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持つ 2 語の場合，以下の構造のいずれかに分けられる（**図 1**）．

(1) 自立語を介さず直接あるいは助詞など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

☆ したがって，本論文では係り受け構造によって表される 2 語の意味の結び付きを意味的関係と呼ぶ．

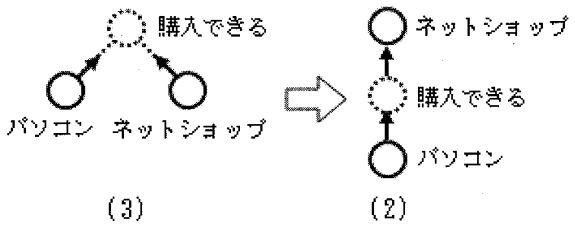


図 2 2語の再構成
Fig. 2 Recomposition of dependency structure.

を除き，一方の語がヘッド☆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．たとえば「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．したがって，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・類義語の関係にない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，さらに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☆☆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，そうでなくても前後の文脈の中に検索対象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が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見なし，その2語が修飾–被修飾の関係を持ち，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．

2.1.2 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，本手法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(A)で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(B)では，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．

(A) クエリの分析

我々は，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1,126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☆☆☆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

☆ 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．

☆☆ 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．

☆☆☆ 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，まった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）と希である．

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．

(B) 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) 「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ないよう 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) で収集されたすべ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．

(1) 「キーワードを 2 つとも含む文」が文書内に存在しない．

(2) 「キーワードを 2 つとも含む文」が文書内に存在する．

(2-1) 2 つのキーワードが適切な修飾–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．

(2-2) 2 つのキーワードが適切な修飾–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) が全体の約 50%，(2-1-1) が約 25%，(2-1-2) が約 5%，(2-2) が約 20%を占めるという結果になった．

さらに (2-1-1) については，主に2つの傾向が見受けられた．1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ーンである．たとえば「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（3.2 節参照）．もう1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ーンである．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．これについては本論文では未対応である．

本研究では，キーワード間に修飾–被修飾関係があるものを判定するための仕組みを作り，(1) と (2-2)，(2-1-1) の (a) のパターンを排除する．よって7割以上の不適合ページが排除できるものと期待できる．

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．再現率に関しては，本論文では係り受け構造を用いて2語の意味的関係を表現している文書しか抽出していないため，現段階では，ある程度の低下はやむをえない．我々の調査では，係り受け構造以外の構造（表構造や見出し構造など）で2語の意味的関係を表現している適合文書が全適合ページ中の約4割を占め，それらの構造は，本論文の検討の範囲外にある．これらの構造への対処は別途報告する．本論文では，適合文書の残り6割について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．本手法の範囲内での再現率についての詳細は 6.5.2 項で論じる．

2.2 関連研究

修飾–被修飾関係という係り受け構造を手がかりとして文書検索を行う研究としては，文献 1)〜6) など数多く存在する．しかし，これらはすべて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[1),2)] は連体助詞句をともなう名詞句（「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ともなって連体修飾節を構成して他方を修飾するパターン（「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[5),6)] 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見なす，

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語を結び付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ー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．さらに，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いまだ検討されていない．

3. 検索性能の改善を目的とした修飾–被修飾関係の詳細な分析

3.1 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などを介して修飾–被修飾関係を持つ際の係り受け構造は，2語の品詞の組合せによって異なる．そこで，2語の品詞の組合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．

3.1.1 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語の組合せについて検討し，次いで個々の組合せごとに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，さらに，2語の意味の組合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．たとえば，「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）」から構成されるととらえ，それぞれが「属性」を持ち，個別の「属性値」をと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（3,776 m，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，その組合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[9]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[10]．たとえば，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[11]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持つ物体．… 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合せとそうでない組合せの弁別をより効果的に行えると考える．たとえば，従来の分類では，「1 kg」「重さ」はいずれも抽象名詞にあたるため「カメラ」と「1 kg」という組合せと「カメラ」と「重さ」という組合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，「1 kg」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合せの場合は「値名詞の実体名詞（1 kg のカメラ）」，実体名詞と属性名詞の組合せの場合は「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 1 kg）」「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☆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

☆ http://www2.nict.go.jp/r/r312/EDR/J_index.html

表 1 概念辞書の対応
Table 1 Matching of our semantic categories and EDR's concept dictionary.

EDR	本手法
1 人間または人間と似た振舞いをする主体	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

およそ**表 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．たとえば，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ともに，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 分類のレベルで語の組合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 種類の名詞の組合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（2.1 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 つはともなうケースが全体の 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．たとえば「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あわせて指定する必要がある．したがって，現象名詞は実体名詞をともな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．実体名詞とともに指定された場合，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ともに 2 語で用いられる場合は，指定された属性値を持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 3,776 m である山を知りたい場合，「高さ 3,776 m」よりも「3,776 m 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）実体名詞とともに用いられて，指定された属性値を持つ実体を検索するときに用いられる．

そこで，名詞 2 語の組合せを

- 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ーンを検討し，整理することにする☆．結果は後出の**表 2** にまとめる．以下の検討における (a)～(f) は，各々表 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」をともな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，この組合せの場合には，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．

この組合せにおいて，実体名詞は現象名詞を用言化した場合の格名詞☆☆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

☆ 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，この場合は実体をともなわず，たとえば「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」の組合せも検討する必要がある．

☆☆ 格助詞を介して述語に係る名詞を「格名詞」と呼ぶ．

表 2 名詞キーワード間の係り受けパターン
Table 2 Dependency patterns between two keyword nouns.

• 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)

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☆場合（「加湿 エアコン」など）(a) (b) がある．

B) 実体名詞＋属性名詞

この組合せにおいて，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持つ．たとえば「スイカ」と「糖度」ならば，「スイカが糖度を内包する」という関係を持つ．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．たとえば，「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（たとえば「スイカが持つ糖度」「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，たとえば「(特定あるいは不定の) 糖度を持つスイカ」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いい難い．そこで，この組合せの場合，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) としてとらえ，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（上述の例でいえば「が持っている」「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．したがって「スイカが持っている糖度」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（たとえば，「スイカは 10～13 程度の糖度を持つ」など）(d)．

C) 実体名詞＋値名詞

この組合せの場合，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，さらに属性と値とが「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．したがって，「値」に等しい属性を内包している「実体」（たとえば「3,000 m 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（上述の例でいえば「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できる）(a) (b)．

☆ 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）．

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つの述語および属性名詞を介して接続している．したがって，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☆．また，1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3,776 m である」など）は考えられるため，このパターンも追加しておく (d)．

D) 実体名詞＋実体名詞

2語ともに実体の場合は，2実体を結び付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．したがって一方の実体名詞は用言ととも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つの実体をあ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ったりする（「万年筆インク」など）などの文構造がと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ーンをそれぞれ考える必要がある．

E) 対象とする係り受けパターン

2語の意味の組合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ーンの基本であるが，それにあてはまらない例外的なパターンも存在する．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ー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付き，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．たとえば，「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とらえることができる．したがってこの場合，実体名詞は現象名詞に係ると見なすことができる．

(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．たとえば「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が係りうる場合であっても，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に係ると見なせる．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語を挟むパターンを加える．これは，たとえば「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ー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2に示すパターンを抽出して解析対象とする．

3.1.2 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合せを考える．

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）の語の組合せだけに限定する．このうち，副詞（および，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（および，形容詞連用形，形容動詞連用形）の組を除

☆「高さが 3,000 m 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，現在のこのパター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．

表 3 名詞以外のキーワードを含む場合の係り受けパターン
Table 3 Dependency patterns between two non_noun keywords.

● 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．

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．したがって，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ー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合せパターンとしては，表 2 で「実体名詞＋現象名詞」の組合せで列挙したパターンのうちで，(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 つのパターンと，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ーン，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ーンとなる．

以上を整理し，**表 3** に名詞以外のキーワードを含む場合の係り受けパターンを示す．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．

3.2 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．したがって，該当する係り受けパター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ー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ー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 2.1.1 項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ー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 つのキーワードのうち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ー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 語の係り受けパター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，4.2 節および 5.2 節で詳述する．

4. 係り受け解析の工夫

3 章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[12)~14)]．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どうしが係り受け関係にある（と人間が判断できる）文をページごとに 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[15)] を利用した．実験の結果，精度はそれぞれ 75%，77%と

低いことが分かった．本来 CaboCha の係り受け正解率☆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．

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．

4.1 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分かった．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）がよく用いられる傾向にある（**表 4**）．

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上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，括弧内のフレーズが括弧外と構文的なつながりを持たないことが多い（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つながりはなくて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見なすべきであり，本論文では扱わない．

☆ 文末の 1 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 用途ごとの括弧記号の使用率
Table 4 Major usage of brackets and frequency of used symbols.

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 1,190 文

n=1,327（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

4.2 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–被修飾関係を持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おさえる．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．したがって，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．たとえば「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持つ語（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，係り受けパターンとの比較処理を行う．接続助詞より文頭側の文に，係り受けパター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．

4.3 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．たとえば，ホテルの予約を行うサイトにおいて，検索されやす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あげられる．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(ただし括弧内を除く)でキーセンテンスを分割し,分割後のセグメントも文と見なし,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.ただし列挙の1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1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4.4節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,たとえば,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.

4.4 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ー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,並列連体助詞および記号(中黒点,カンマ,読点,アンパサンド)と定める.

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

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見なす.

以上の処理によって,たとえば,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.

4.5 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.たとえば,「世界の電圧>東南アジア>マレーシア」といった表記は,文とはいい難いが,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「<」がないにもかかわらず「>」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.

5. システム構成

本手法のアルゴリズムを,既存の検索エンジンの検索結果を並べ替えるフィルタリングツールとして構築する.

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☆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.

5.1 システム全体の流れ

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,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.このカテゴリを用いて係り受けパターンの候補をあげる.

それと並行して,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.各ページに対して,次節で述べるアルゴリズムに基づき,係り受けパターンに該当するか否かを調べる.

そして既存の検索エンジンによる順位出力は本システムによる判定の結果によって並べ替える.すなわち,修飾–被修飾関係にあると判定された(パターンにマッチした)か否かによって2分割し,それぞれのグループ内でもとの順位関係を保ったまま,修飾–被修飾関係にあるグループを上位に,ないグループを下位にランキングする.並べ替えられた検索結果はユーザに提示される.

☆ Google APIs http://www.google.com/apis/

5.2 文の係り受け構造を用いた判定の方法

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．

(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．ただし，ピリオドの前後が数字であった場合は，小数点と見なし，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 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 4.2 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 4.3 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ーンを比較する．キーセンテンスの一部がこれらのパター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，係り受けパターンと一致したと見なす．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ー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ー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（3.2 節参照）．

(9) (7) においてパターンと一致しなかった場合は，並列構造解析を行う（詳細は 4.4 節参照）．

(10) 文中に「<」がなく「>」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見なし，「>」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 4.5 節参照）．

6. 評価実験

6.1 評価データの作成

オープンテストにより本手法を評価するため，以下の (1)～(3) の手順で設定した検索キーワード対について Google を用いて検索を行い，上位 100 件までにランキングされたページをもとに，データセットを作成する．

(1) ウェブページを 10 万ページ取得して形態素解析を行う．

(2) 一般的すぎる語では漠然としていて検索キーワードとしてふさわしくないため，TF-IDF で上位となった語から，人手で検索キーワードとして妥当な組合せを 2,000 組作る．その際，漠然とした意味の（抽象度の高い）語ばかりだと検索意図が推定しやすいクエリが作りにくかったことから，共起確率の高い語（同一文書に同時に出現しやすい語）どうし（たとえば「サッカー」と「チーム」など）について複合語とすることも認める．

(3) (2) の作業にかかわらない複数人（今回は 3 人）が同一の検索意図を推定できるものを選ぶ．

以上の手順で決定した 118 組の 2 語キーワード（実体名詞どうしの組合せが 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）から，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およびバイナリファイルを除いた）11,776 ページについて，適合/不適合の判定を行った．判定は，3 人の学生によって行い，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．

6.2 係り受けパターンの妥当性の検証

表 2，表 3 にあげた係り受けパターンの妥当性を検証する．この検証にあたり，6.1 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．11,776 ページから，抽出されたキーセンテンス計 12,536 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** に示す．表中の列見出しの「A」は A が

表 5 係りのタイプ別精度
Table 5 Accuracy of each type of modification.

n=12,536（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	1,836	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	2,203		310		91		222	0	552		0	0
不適合	1,226		82		19		73	0	155		0	2
精度	0.642		0.791		0.827		0.753		0.781			0.000

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分かる．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．

6.3 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ーンを用いた手法との比較

6.3.1 評価手法

比較の評価指標には，上位 100 件中の精度・再現率・F 値・MAP を用いる．値を算出するために，各手法ごとの基準に基づき，100 件のデータを「優先」と「非優先」の 2 グループに分ける．精度・再現率・F 値は,「優先」グループを適合として算出する．一方 MAP はランキングの良し悪しを評価するための指標である．そこで「優先」グループを，もとの検索エンジンの順位に従って並べ，その後に「非優先」グループをもとの検索エンジンの順位ごとに並べるという手法に基づきランキングを行う．MAP（Mean Average Precision）とは，各検索課題ごとの平均精度の平均である．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．

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

データセットは 6.1 節のデータセットから 20 組（**表 6**）を抽出し，これを利用する．20 組の選定にあたっては，実体＋値の組合せは他の組合せと比較して，現実の検索で用いられる可能性が低く，また実体＋属性の組合せも実体＋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[16] 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．

6.3.2 各手法との比較

以下にあげた (i)～(v) の各手法においての実験結果を**表 7** に示す．なお本項では，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「係り受けパター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持つ文を含む文書を優先させる．

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

		検索キーワード
(1)	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

表 7 提案手法と類似手法の精度・再現率・F 値・MAP
Table 7 Precision, recall, F-value and MAP of our method and similar methods.

n=1,995（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

(ii) 表 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 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どうし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 7 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 7 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，6.5.2 項で考察する．

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☆．

続いて (iv) は，表 2・表 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ーンを用いた手法（表 7 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 がセンテンス単位の集計であったのに対し，表 7 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，6.2 節で行った網掛け部分についての考察のとおり，ページ単位で集計することによって，表 7 において差は 1.0%に広がった．

(iii) は表 2 に 5.2 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☆☆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，2.2 節であげた関連研究は，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．

6.4 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[17]．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やすい，ランキング上位部分に適合文書が並ぶこと

☆ 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．

☆☆ 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．

表 8 提案手法と類似手法の適合ページ数と平均精度
Table 8 The number of relevant pages and MAP of our method and similar methods.

	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

が望ましい[18]．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，6.1 節で述べた，検索結果上位 100 件に対して，我々のシステム（6.3 節 (iii) に 3.2 節「係り受けパター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[19] 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 8 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 7 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 7 および表 8 のそれぞれの手法は，対象データセットおよ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 8 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．

6.5 考　　察

6.5.1 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.

6.5.2 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7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見たところ,修飾–被修飾関係にな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).たとえば,「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いえることは,「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えない」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7のいずれの方法でも抽出できていない.したがって,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.

再現率を低下させるもう1つの主要な要因として,3.1節で検討したパターン以外の構文パターンで2語の関係が表されている文の存在をあげることができる(約4割).たとえば,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,ある程度パター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.

6.5.3 文中における位置の戦略に対する効果

表7および表8のすべてのMAPを比較すると,表8(D)のMAPが最も高く0.739であるものの,表7(iii)のMAPは0.738であり,戦略の違いである「係りパターンの文中における位置」が有意といえる差はない.

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分かった.したがって,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.たとえば,「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かかわらず係り受けパター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.

6.5.4 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「7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.6.3節の実験に用いたデータ1,995ページ中,不適合ページは962ページ含まれている.表7(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–被修飾関係にないと判断している.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.

6.5.5 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語でおおむ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,6.1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

表 9 提案手法によるリランキング前後の適合ページ数
Table 9 The number of relevant pages by our reranking method.

	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

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 9** 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いいがたいケース（たとえば(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．

7. おわりに

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など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．検索キーワード間の修飾–被修飾関係を利用したウェブ検索エンジンの精度の向上について議論を行った．文の中に現れる修飾–被修飾関係を確認する手法として，検索キーワードの意味分類に応じた加味すべき係り受けパターンを整理した．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．これを用いて係り受けパターンの妥当性を示し，また係り受けを用いた他の手法との比較実験を行った．フィルタリングツールとしての評価においては，我々のシステムがもとにした検索エンジンおよび近接距離に基づいた手法の精度を，上位20位において7～10%程度向上させることを示した．

本手法は，整理されたパターンの係り受け構造を持つ文を含む文書を抽出するという方法であるため，再現率の点ではキーセンテンスを抽出する方法よりも下回ったが，表構造/見出し構造を利用することで，それと同等以上にすることが可能である．さらに，フィルタリング結果の提示方法，スコアリング方法を工夫することで，検索エンジンとしてのユーザ満足度を向上させることも可能である．

今後は，表構造/見出し構造に着目したシステムの拡張を行うとともに，フィルタリング結果の提示やスコアリングの戦略についても検討を進める予定である．

参考文献

1) Mine, T., Fujitani, H. and Amamiya, M.: A Japanese Information Retrieval Method Using Syntactic and Statistical Information, *NLPRS2001*, pp.429–434 (2001).

2) 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).
3) Strzalkowski, T. and Carballo, J.P.: Recent Developments Natural Language Text Retrieval, *The 2nd Text REtrieval Confernce (TREC-2)*, pp.123–136, NIST Special Publication (1993).
4) Zhai, C., Tong, X., Milic-Frayling, N. and Evans, D.A.: Evaluation of Syntactic Phrase Indexing — CLARIT NLP Track Report, *The 5th Text REtrieval Confernce (TREC-5)*, pp.347–358, NIST Special Publication (1996).
5) 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—ダイアログナビ，自然言語処理，Vol.10, No.4, pp.145–175 (2003).
6) 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).
7) Jansen, B., Spink, A. and Saracevic, T.: 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).
8) 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).
9) Takagi, A., Asoh, H., Itoh, Y., Kondo, M., and Kobayashi, I.: 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).
10) 亀井 孝，河野六郎，千野栄一：言語学大辞典，Vol.6 術語編，三省堂 (1996).
11) 斎藤秀三郎，松田福松：名詞用法詳解，吾妻書房 (1956).
12) 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).
13) 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).
14) 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).
15) 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).
16) Matsumoto, A., Konish, T., Takagi, A., Koyama, T., Miyake, Y. and Itoh, Y.: 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).
17) Spink, A., Jansen, B.J., Wolfram, D. and Saracevic, T.: From E-Sex to E-Commerce: Web Search Changes, *IEEE Comput.*, Vol.35, No.3, pp.107–109 (2002).
18) 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).
19) 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).

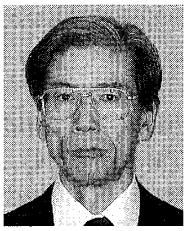
(平成 18 年 9 月 1 日受付)
(平成 19 年 7 月 3 日採録)

松本 章代（正会員）

2004 年静岡大学大学院情報学研究科修士課程修了．同大学院理工学研究科博士後期課程在学中．2005 年東京工業高等専門学校助手．現在に至る．自然言語処理，情報検索に興味を持つ．電子情報通信学会会員．

小西 達裕（正会員）

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



高木　朗（正会員）

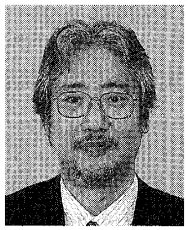
1974年早稲田大学大学院理工学研究科修士課程修了．1974年早稲田大学大学院博士後期課程編入．1981年早稲田大学大学院理工学研究科研究生．1983年（株）CSK（現，（株）CSKシステムズ）入社．2007年2月退社．言語情報処理研究所設立．現在に至る．（独）産業技術総合研究所客員研究員．自然言語処理等に関心を持つ．工学博士．電子情報通信学会，人工知能学会，日本認知科学会各会員．

小山　照夫（正会員）

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

三宅　芳雄（正会員）

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



伊東　幸宏（正会員）

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