単語間の係受け関係を用いた情報検索手法の評価

松村 敦[†] 高須淳宏[†] 安達 淳[†]

従来から行なわれている文書検索手法は，出現単語とその統計的特徴量のみに頼ったものであるため，その精度には限界がある．そこで我々は，単語間の係受け関係を考慮する検索手法を開発し，検索精度を高めることを目指す．本手法では，論文タイトルに含まれる単語間の係受け関係を二分木の形に構造化しインデクスに取り込み，同様に，タイトルのような疑似的な自然文による問合せも構造化する．これら二つの二分木間の類似度を定義し，検索結果のランク付けを行なう．検索実験による評価を行なったところ，IDF 法に比べて検索精度の向上が見られた．

Evaluation of Information Retrieval Method Using Dependency Relationships Between Words

Atsushi Matsumura,[†] Atsuhiro Takasu[†] and Jun Adachi[†]

Although keyword-based retrieval occupies the major position in document retrieval methods, it has obviously limitation in retrieval effectiveness because it is difficult to satisfy users' requests using only query words and its statistical characteristics. Thus we have developed a method to enhance retrieval effectiveness using dependency relationships between words in sentences of documents and queries. In our method, we make the dependency relationships between words in titles of documents and queries into structured index represented with binary trees. Rankings of all retrieved documents are calculated using the semantic similarity between the two binary trees. Through the retrieval experiments, we show that our method is more effective than IDF method.

1. はじめに

近年，文書の電子化が進み，大量の電子情報が蓄積されるようになってきた．また，計算機やインターネットの普及によってこれらの電子化された情報へのアクセスが多くの人達に解放されつつある．このような環境では，大量の情報の中から欲しい情報だけを効率良く抽出する技術が重要である．従来，情報検索の研究は所望の文献を探し出すことが中心的な課題であり，その主流となっている手法はキーワードを用いた論理検索である．しかしながらキーワード検索では，検索意図を論理式に変換した時点でキーワード同士の意味的関係が失われるので問合せ自体に曖昧性が残ってしまう，といった問題が生じる．

こういった問題の解決のために，ベクトル空間モデルや TF-IDF （Term Frequency Inverted Document Frequency）法などを用いた得点付け[1]を行う検索エンジンもある[2]．しかしながら，これらの手法は単語の出現頻度等の統計的特徴量のみに依存し，キーワード間の意味的な関係を補うことができないという限界がある．

一方で文の意味を理解することにより，情報検索の精度を向上させようという試みが，自然言語処理の研究を通して行なわれている[3]．しかし，このような高度で複雑な処理を行うとシステムの負荷も大きく，また，高い検索精度を達成する場合には適用の場を限定する必要がある[4),5]．したがって，より広範に適用可能な実用的システムの構築を目指す場合には，できるだけ複雑な処理を使わずに文書や問合せに意味内容を付与することを考えなければならない．

本論文で我々は，比較的処理の負荷が小さく精度も高い形態素解析のみを使って，単語間の係受け関係を検索に利用することを提案する．単語の頻度情報以外に，単語間の係受けという意味情報を文書，および問合せに与えることで検索精度の向上を目指す．

単語間の意味的関係を利用した情報検索の研究は，これまでにも盛んに行われてきた．単語間の統語的な関係を用いた情報検索の研究[6),7]では，4つの統語的関係を利用して情報検索に利用しているが，十分な精

† 学術情報センター
National Center for Science Information Systems

度向上は得られていない．また，単語間の意味的な関係を利用して，文書と問合せを構造化した研究が Lu によって行われ[8]，情報検索への応用の可能性が示唆されている．しかしながら，構造化の過程が手作業であり，また十分な評価実験が行われているとは言い難い．一方，Myaeng らは自然言語処理を使って文書や問合せを Conceptual Graph 表現に置き換え，情報検索に応用するための研究をすすめているが[9]，自然言語解析の評価が主であり情報検索への応用は試験段階である．このように英語圏では，単語間の意味情報まで利用した情報検索の研究が進められているが，日本語における研究は遅れている．

本論文では，単語間の意味的な関係として係受け関係に着目し，これを用いた情報検索手法の日本語文書への適用を提案しその効果を議論する．以下，2 章では我々の手法である「構造化インデクス」について概要を説明し，3 章でインデクス化の手法，4 章で検索の手法について解説する．検索のための評価実験について 5 章で述べ，6 章で結論と今後の課題を述べる．

2. 構造化インデクス検索システム

我々はこれまで，プロトタイプシステムを使って単語間の係受け関係を利用した検索手法の有効性を探ってきた[10),11)]．その結果，従来から使われてきた TF–IDF 法と比較して性能向上を実現できることが分かった[12)~14)]．ここでは，我々の提案する係受け関係を利用した情報検索の枠組について説明する．

文に現れる二つの単語，または単語のまとまりと，それらを関係付ける語を抽出し，これを二分木の形に構造化し文の内部表現として扱う．この二分木を検索インデクスとして利用することによって検索精度の向上を目指す．本研究では，このように構造化されたインデクスを「構造化インデクス」と呼んでいる．構造化インデクスの一例を図 **1** に示す．構造化インデクスは，キーワードおよびそれらの関係を表す語から成り立っている．

我々の枠組において，キーワードはある概念を表すため，「概念語」と呼ぶことにする．文法要素としては，「名詞」「形容詞」「副詞」がこれにあたる．また，複合名詞はその構成単語にまで分割することができるので，概念語の並びと考える．一方，概念語どうしの関係を表す語を「関係語」と呼ぶ．関係語は文法要素で言うならば「助詞」「助動詞」「動詞」及び，それらの組合せなどである．関係語は「に関する」「についての」のようにほぼ同じ意味を有するものを一まとめにし分類することができるが，これをカテゴリと呼ぶ．

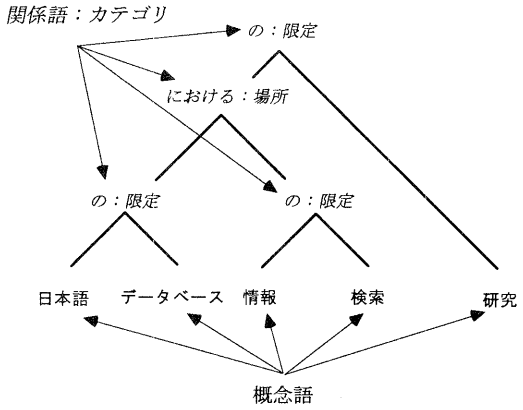

図 1 構造化インデクスの例
Fig. 1 Sample of a Structured Index

構造化インデクスは図 1 にあるように，二分木の葉の部分に文を構成している概念語，内部ノードに概念語同士を結び係受け関係を表す関係語及びそのカテゴリを配するデータ構造を持つ．

しかしながら，自然文に対してこのような係受け関係を解析することは複雑な自然言語処理を必要とし，一般的にはシステムの負荷が大きく精度も低いため，実用的なシステムを作成するのは難しい．そこで，本検索システムは文書のタイトルのような疑似的な自然文を検索対象とする．これは

- 自然文に比べて構造が単純．
- 文書の主題を端的に表しており，かつ抽出が容易．

という特徴を持つ．構造の単純さゆえに係受けの付与が自然文と比べると容易であり，以下で我々が提案するタイトルパタンによる解析が精度良く行われ，システムの負荷も少ない．また文書の主題を端的に表すため文書の持つ情報を代表させるのにも相応しいと考えられる．最近，同様な試みが韓国語においてなされているが，情報検索への応用には至っていない[15)]．

提案システムの全体像は，図 **2** のように示すことができる．文書から抽出されたタイトルは，形態素解析，係受け解析，複合名詞解析の三段階の処理を経て，構造化インデクスとしてデータベースに蓄えられる．一方，疑似的な自然文として発せられた問合せも同様の過程を経て，構造化される．これら二分木どうしのマッチングを取り意味の類似性にしたがって文書に得点が与えられ，順位付けされた結果が出力される．

3. 構造化インデクスの作成

3.1 形態素解析

文を概念語と関係語に分類するために，はじめに形態素解析によって文法要素に区切る．ここで，複合名

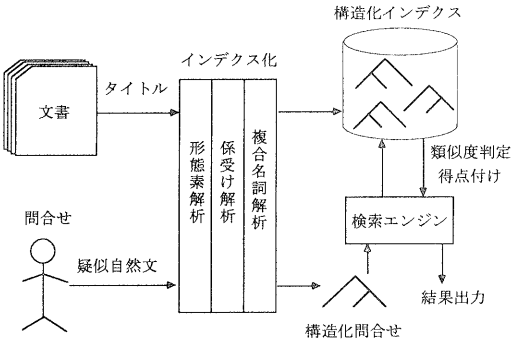

図 2 提案システムの処理の流れ
Fig. 2 Processing flow of our system

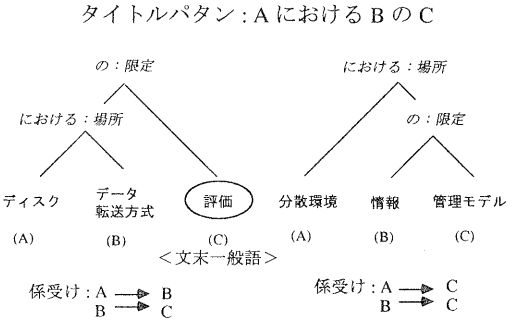

図 3 係受けの曖昧性
Fig. 3 Ambiguity of the dependency

表 1 関係語のカテゴリとその主な要素
Table 1 All categories of relation words and their typical elements.

カテゴリ名	主な関係語
限定	の，な，された，される
場所	における，での，上の，から見た
手段	による，を用いた，に基づく，を利用した
and	と，および，ならびに，も
目的	のための，を目指した，を指向した
内容	に関する，についての
方向	への，向きの
起点	からの，から
考慮	を考慮した，に着目した，を想定した
対象	に対する，を対象とした
所有	を持つ，を有する，を持った，を備えた
共有	間の，で共有された，間での
同格	としての
補助	を支援する，をサポートした
主格	が，は
適応	に対応した，に適した，に応じた
可能	可能な，を可能とする，が可能な
or	や

詞はその構成概念語にまで分割される．形態素解析ツールは茶筌[16]を利用した．次に，形態素解析の結果を元に，品詞情報と関係語辞書を用いて概念語と関係語を同定する．ここで，活用語尾の変化で吸収されるものに関しては，同一の関係語としてシステムが自動的に認識する．関係語辞書は，あらかじめ 3666 件の科学技術系論文のタイトルから人手で関係語を抽出し作成した．抽出した関係語は 256 種類 6763 個である．これらの関係語は，意味の類似性によってカテゴリ分けされる．本システムでは，18 のカテゴリを定義したところ，約 98 % の関係語がそれらに分類された．どのカテゴリにも属さない残りの関係語は「その他」というカテゴリを作成して分類する．**表 1** に 18 のカテゴリとそこに分類される主な関係語の例を示す．

3.2 係受け解析

次に，分割された概念語や複合名詞の間に係受け関係を付与する．本研究では「タイトルパタン」と係受け構造の対応表によって係受け関係を同定する．タイトルパタンとは文中の関係語の並びのことである．例えば「係受け情報を用いた文献検索システムの評価」という文は A，B，C を概念語またはその並びである複合名詞として，「A を用いた B の C」というタイトルパタンに属するという．この対応表はあらかじめタイトルを分析し作成した．分析対象は関係語を抽出した科学技術系論文のタイトル 3666 件である．このうち名詞数が 3 又は 4 のタイトルを調べ，タイトルパタンとして頻度 3 以上のもの 105 種を抽出した．これらのタイトルパタンに対して人手で係受けを付与し，タイトルパタンと係受け構造の対応表を作成した．

その結果，殆んどのタイトルに唯一の係受けが付与される確定的な場合と，タイトルに応じて複数種類の係受けが付与される曖昧な場合があることが分かった．後者は例えば「A における B の C」という場合で，二つの係受け構造をとる可能性がある（**図 3**）．このように係受け構造に 2 種類の選択肢があるタイトルパタンではパタンの最後の単語の性質を調べることによって，高確率で正しい係受け構造を同定できることが分かった．判断の基準は，タイトルパタンの末尾の単語が「研究」「提案」などタイトル全体の特徴を表し，タイトル中の特定の単語との係受けにあまり関与せず，前全体を受ける係受けを有する単語かどうかである．このような性質の単語を文末一般語と呼ぶことにし，本システムでは上記の語の他「評価」「実現」等 53 語を定義した．図 3 では，左の場合は「評価」という文末一般語があるため「A → B」および「B → C」という係受け構造となる．一方，文末一般語がない右の場合には「A → C」および「B → C」という係受け

表 2 タイトルパタン，拡張タイトルパタンの数
Table 2 The number of Title Patterns and Extended Title Patterns

名詞数	タイトルパタン数	拡張タイトルパタン数
3	62	40
4	43	33

構造となる．

しかしながら，関係語の多様性からタイトルパタンで全てのタイトルに係受けを付与することができるわけではない．そこで，タイトルパタンの概念を関係語のカテゴリに適用したものに拡張する．例えば，「A から見た B に関する C」というタイトルの場合，同じパタンのタイトルが少く，対応表には登録されていない．この場合「A[場所]B[内容]C」のようにタイトルパタンを拡張することによって，係受け構造が付与できる．このようなパタンを「拡張タイトルパタン」と呼ぶ．本研究で使用したタイトルパタン，拡張タイトルパタンの数を**表 2** にまとめた．

5 名詞以上のタイトルの場合や，一部の 4 名詞の場合で拡張タイトルパタンを用いても係受けが判定できない場合には，いくつかの経験則にしたがって単語数を減らし係受けを付与する．このような方式をとると，文全体に正しい係受けが付与されなかった場合にも局所的には正しい係受けが付与される可能性が高く，係受け付与のミスによる検索精度の劣化が押えられる．

3.3 複合名詞解析

タイトルパタンによる係受け解析では一まとまりとして扱った複合名詞も，その構成単語間の係受けを考えることによって構造化インデクスの枠組に組み込むことができる．これによって，複合名詞の意味を反映した検索が可能になり，検索精度の劣化を押えることができる．

複合名詞は構成する単語間に適当な付属語を補いながら読み換えることで通常の文に変換できることが知られている[17)18)]．本研究ではこの点に着目して，複合名詞を適当な自然文に読みかえて，分割された複合名詞の構成単語間の係受けを定め，構造化インデクスの枠組へ組み込むことを考える．

単語間に補う関係語については，「化」「的」のような接尾辞で終っている場合には，「化された」「的な」のように特徴的な関係語を補って読み換える．接尾辞がない場合には，基本的に「の」という関係語を補うこととする．これは，予備的な実験によって単語と関係語の共起関係を調べたところ，「の」という関係語との共起関係が最も多く，統計的に「の」で補うのがもっとも正確であると考えられるからである．名詞の語彙的情報を用いて詳しく関係語を規定することも可能だが，辞書の保守が容易でないなど問題が多いため，ここでは表層情報から得られる知識のみでの解析を試みる．

構成単語間の係受け解析においては，意味的結合関係の強さの指標として単語 bigram 統計を利用することを提案する．単語 bigram とは，複合名詞中の 2 個の連続した単語の組合せがどのような頻度で生じるか統計を得たものであり，十分な標本数を取ればそのまま単語間の意味的結合関係の強さを表すことができると考えられる．したがって，単語 bigram の大きい単語の組に優先的に係受けの付与を行う．本研究では，13615 件の論文タイトルから 60507 個の単語 bigram を自動抽出した．このうち頻度が 10 以上のもの 814 種類を単語 bigram データとして利用した．

4. 検　索

4.1 問合せの構造化

本システムではタイトルのような疑似的な自然文を問合せとして採用し，前節と同様の手順で係受け関係を表現した二分木に構造化する．その結果，問合せに含まれるキーワード間の意味的な関係を扱うことが可能となり，問合せの過程で生じる曖昧性が緩和される．

このような問合せの処理方式を採用することにより，本システムでの検索とは係受け関係を表現した二分木間のマッチングをとることに帰結される．以下では，本システムで採用した係受け関係を表す二分木間の意味的な一致度の定義とマッチングのとり方，得点付けについて述べる．

4.2 単語のマッチングと得点

問合せに含まれる単語が，タイトル中に含まれているかどうかは重要な要素であり，本手法でもこの要素をまず考慮に入れる．単語のマッチングについては，頻度情報を考慮に入れた重要度を計算して一致度を定義する TF-IDF 法が一般的である．本手法においても，単語のマッチングは TF-IDF 法に従う．ただし，今回扱うタイトルのような文は高々名詞数個の文であり TF はほぼ 1 として無視できる．したがって，単語に対する得点は IDF のみを考慮した式 (1) によって算出する．

$$S_b(W) = \frac{\sum_{i=1}^{n} \log \frac{S_{all}}{S_i} \cdot \delta_i}{\sum_{i=1}^{n} \log \frac{S_{all}}{S_i}} \qquad (1)$$

ただし，式中の変数はそれぞれ以下の通りである．

$S_b(W)$：文書 b における単語点
i：タイトル中の単語の番号
n：タイトル中の単語の総数
δ_i：$\begin{cases} 1 & \text{単語 } i \text{ がマッチした場合} \\ 0 & \text{otherwise} \end{cases}$
S_{all}：検索対象文書の総数
S_i：単語 i が出現する検索対象文書数

また，式（1）は取り得る最高得点で正規化されており，1点満点となる．

4.3 係受け関係のマッチングと得点

4.3.1 二分木の縮退

係受け関係を用いた検索では，問合せと一致する係受け関係を含む検索対象文が適合文候補となる．本研究では，はじめに問合せに含まれる係る語，係られる語を全て含む検索対象文を候補として抽出する．次に，候補の係受け関係を表わす二分木の構造が問合せの二分木を包含するかどうかを判定する．包含した場合には，係受け関係に対して重要度を判定し，それに応じた得点を与えるという手順を踏む．

ここで，単純に問合せの二分木を包含する二分木だけを抽出すると，問合せに含まれる情報のうちの1つでも情報の欠落した文書，あるいは問合せに含まれる係受け関係の間に余分に単語が挿入されている文書は検索されないという問題が生じ，検索漏れの原因となる．そこで日本語の係受けの性質に則った木構造の縮退動作を行うことでより柔軟な検索を行うことを提案する．ここでいう二分木の縮退とは，文中の係受けを部分的に消去することで文意を極力変化させずに二分木を縮小する動作のことを言う．基本規則は，あるノードを縮退させる場合には右側の子孫を残すということである．具体的な縮退の例を図4に示す．縮退操作は問合せとタイトルの双方に適用され，両者の二分木に包含関係が生じるまで繰り返される．

4.3.2 係受けの重要度と得点

二分木の縮退操作によって，問合せに含まれる係受け関係を包含する検索対象文が検索結果の候補として特定される．次に，これらの係受け関係に対して，意味的な重要度に応じた得点を与えることを考える．本研究で採用した係受けの重要度は以下の3種類である．

意味的な類似性 二つの係受けが意味的に類似しているかどうかは，検索する上で重要な要素である．同じ係る語と係られる語を持っている場合に，本研究では次の3つのレベルを設定し係受けの意味的類似性を評価する．

Exact Match 係受け関係を表す関係語が等しい．

Category Match 係受け関係を表す関係語は異なるが，同じカテゴリに属する．

Wild Match 係受け関係を表す関係語もその属するカテゴリも異なる．

それぞれのレベルに応じて得点を与えることによって，意味の違いを意識した検索が可能となる．

検索対象データベース中での重要度 検索する係受け関係が検索対象データベース中でどの程度重要な概念であるかを評価する．本システムでは係受け関係を構成する係る語，係られる語のデータベース中での重要度をその係受けの重要度と置き換える．一般に，重要な単語に対する係受けほど重要と考えられるからである．単語の重要度は，データベースにおける単語の頻度と一般語という特殊な単語を考慮することで評価する．

データベース中で頻度が高い単語はそれほど重要ではなく，出現頻度の少ない単語ほどそのデータベース中では特徴的で，重要な概念に関係していると考えることができる．係る語，係られる語両方の寄与を考慮するため，これら二つの単語の頻度の相乗平均を係受けの重要度として採用する．また，一般的な概念を表す単語である「一般語」が係る語，または係られる語に含まれているかどうかは別途考慮する．何故なら，これらの単語は他の単語に比べてノイズとなる可能性が極めて大きいからである．本手法では，係受け関係を持つ係る語，係られる語に含まれる一般語の数によって3段階で評価をする．

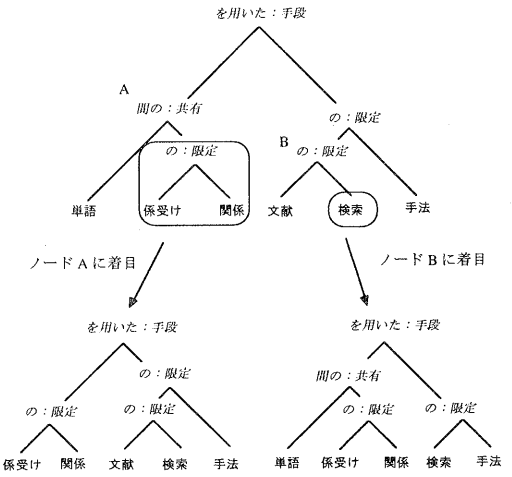

図4 二分木構造の縮退
Fig. 4 Shrinking the binary tree

文中における寄与度 候補のタイトル内において，特定の係受け関係がどれほどの寄与を持っているかは，その係受け関係の表わす内容と候補文の内容の一致度に関わってくる．この一致度の評価の指標として，文中の係受けの深さを採用する．文を階層構造に組み上げることによって，名詞，複合名詞，名詞句というように階層の深さの順番に段階を設けることができるが，これらの文全体に対する寄与はこの順で大きくなる．したがって，文の重要な内容はより深い係受けによって表現されていると考えられ，この深さに応じて，重要度を評価することは妥当だと思われる．

以上によってマッチした係受け関係の重要度を測り，それぞれの寄与を考慮した得点を加える．係受けによる得点は式（2）によって与えられる．

$$S_b(D) = \frac{\sum_{j=1}^{N} \frac{w_j t_j l_j}{k^{d_j}}}{\sum_{j=1}^{N} \frac{t_j l_j}{k^{d_j}}} \quad (2)$$

ただし，l_jは各係受けのデータベース中での重要度であり，

$$l_j = (\log \frac{S_{all}}{f_j})^2 \quad (3)$$

である．なお，式中の変数は以下の通りである．

$S_b(D)$：文書 b における係受けの点数
j：タイトルの二分木におけるノード番号
N：ノードの総数
w_j：マッチングレベルに応じた重み（≤ 1.0）
t_j：一般語の重み（≤ 1.0）
k：階層の深さにおける重み（≥ 1.0）
d_j：階層の深さ（根が 0）
S_{all}：全文書数（＝タイトル総数）
f_j：子ノードの頻度の相乗平均

また，式（2）は取り得る最高得点で正規化されており，1 点満点となる．

4.4 文書の得点とランキング

本手法では，式（1）によって得られた単語得点と式（2）で得られた係受けによる得点の統合方法として，線形結合を採用する．それぞれの得点に重みをつけて線形結合し 100 点満点換算したものを文書の得点とする．これは

$$S_b = \frac{100 \times x_W}{x_W + x_D} S_b(W) + \frac{100 \times x_D}{x_W + x_D} S_b(D) \quad (4)$$

で表される．ここで，x_W及び x_Dはそれぞれ単語点，係受け点の重みであり，$x_W = 1 - x_D (0 \leq x_W \leq 1)$ の関係にある．このような得点付けを採用することによって，本検索システムは，IDF 法（$x_D = 0$）を基礎とした応用手法と位置付けられる．

5. 評価実験

本検索手法の有効性を評価するために検索実験を行った．検索システムの評価における要素は，検索対象データベース，問合せ，それに対する正解集合及び評価指標の 4 つがある．今回の評価実験は以下のような環境で行った．

検索対象データベース 学術情報センターの提供するデータベース NACSIS-IR からコンピュータ関連の論文 13615 件の書誌情報．

問合せ集合 以下の 5 件で行った．

（1）「類似図形パターン検索」
（2）「WWW 情報空間でのユーザ支援」
（3）「ATM 網におけるトラヒック制御」
（4）「ニューラルネットを利用した画像の特徴量抽出」
（5）「並列処理を用いた FFT の高速化」

正解集合 以下の基準で二つの正解ランクを定義し，5 件の問合せについてそれぞれ作成した．

ランク **A** 問合せの内容を殆んど全て含むもの．
ランク **B** 問合せの内容の一部を含むものなど，関連の深いもの．

ここで，ランク A の数は少なく十分な評価が出来ないため，今回の実験ではランク A とランク B の和集合を正解集合として採用した．

評価指標 評価の指標には，順位を考慮した再現率と適合率を用いた．また，従来から使われている IDF 法を比較対象とし，5 件の問合せについての平均を計算した．

実験は，5 件の問合せについてパラメタを振り最も効率のよい条件の探索を試みた．ここで独立に変化するパラメタは式（2）を調節する 6 個と式（4）の 1 個の計 7 個である．はじめに，式（2）の 6 個を適当に定め，式（4）のパラメタの最適値を探ったところ，$x_D = 0.25$，すなわち，係受けの割合が 25%の時にもっとも性能がよいという結果を得た．そこで，次に x_Dを 0.25 に固定し，残りの 6 個のパラメタの最適化を試みた．その結果得られたパラメタを**表 3**，その時の 5 つの問合せの平均での再現率–適合率曲線を本手法と IDF 法とで比較したものを図 **5** に示す．

図 5 において，横軸は再現率，縦軸は適合率であり，グラフはいずれも，5 つの問合せの平均である．これをみると，係受け点が 25%とした本手法による検索は

表 3 本検索実験におけるパラメータ
Table 3 Parameters in the retrieval experiments

パラメータ	設定値
w_j	1.0 (Exact Match)
	0.9 (Category Match)
	0.8 (Wild Match)
k	1.0
t_j	0.0 (片方が一般語)
	0.0 (双方一般語)

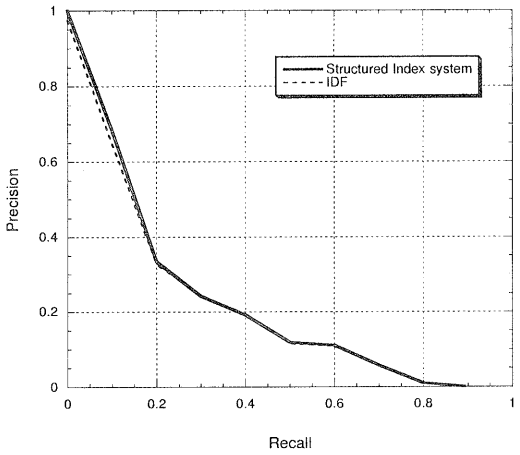

図 5 再現率-適合率曲線
Fig. 5 Average recall-precision curves

再現率が 0.1 までで IDF 法による検索に比べて適合率が約 10%向上していることが分かる．再現率が 0.2 の時まではわずかに本手法の方が適合率がよいが，それ以降は，ほぼ IDF 法による検索と同じ結果になっている．再現率が低いところで適合率の改善が見えることから，順位が高い時の検索の効率は上がっていると考えられる．また，再現率が上がるにしたがって，IDF 法の結果に収束していくのは，係受けのない文書の順位付けは，本手法と IDF 法とで同じになることから予想される結果である．

次に，本手法による検索の効果を見るために，実際の検索における順位付けを分析してみる．**表 4** と**表 5** は問合せ (2) のときの IDF 法と本手法での検索結果の一部である． 表 4 に示されている三つのタイトルはすべて「WWW」と「ユーザ」という二つの単語がマッチしているが，一番上の文書だけは不正解である．IDF 法は単語の頻度のみに頼っているため，このような文書を排除することができない．これに対し，係受け関係を考慮に入れた本手法による結果では，この不正解文書は 26 位にまで順位が下がり，残りの 2 個の正解文書は 4 位に順位を上げている．この例が示すように，係受け関係を考慮することによって，単語の頻度情報のみでは実現できない意味的な情報の抽出が可能となり，結果として検索の精度をあげることができた．

しかしながら，再現率が 0.1 までで 10%程度の精度向上では十分とは言えない．次に，精度が十分に上がらない理由について調べるために，他の問合せの出力結果を考察する．**表 6** は問合せ (3) についての本手法の検索結果の一部である． この場合は，問合せに含まれる「ATM」という特徴的な単語が重要であり，実際に「ATM」という単語を一語含んだだけで正解である文書が多い．しかしながら，検索結果をみると「網における制御」といったそれほど重要でない単語間の係受けを持つ不正解文書の方がより上位に順位付けされるという結果になっている．この例は，語自身の重要性と係受けによる重要性の扱いがまだ十分でないことを示しており，これが検索精度の十分な向上を妨げる一つの原因となっている．

6. 結 論

本論文では，単語間の係受け関係を利用した情報検索手法を提案した．評価実験において，従来から使われている単語の頻度情報のみによる検索手法の IDF 法に比べて，再現率 0.1 までで適合率が 10%程度上回る結果を得ることができた．

検索効率が向上した理由は，同じ単語が一致する複数の文書を係受けの違いによって区別することが可能になったためである．一方，係受け関係の考慮が検索精度を落すこともあることが分かった．その原因の一つとして，係受けの無い重要な単語を含む正解文書よりもあまり重要でない単語間の係受けを含む不正解文書の方が順位が上がってしまい，ノイズが増えてしまうといった現象が見られた．

本手法の目的とするのは，単語の頻度情報による検索では正解か不正解かの区別が出来ない文書を，単語間の係受け関係を表す表現を利用することによって正しく分類し，検索精度をさらにあげることである．今回の結果から今後の課題は二つの段階に分けて考える必要がある．一つは，正しく係受け関係を表す表現が抽出できているかという問題，もう一つは，抽出した係受け関係を情報検索に有効に利用できているかという問題である．

前者の解決のためには，形態素解析，係受け解析，複合名詞解析の精度を向上させる必要があり，その過程では係受けの精度を評価するための正解データの作成も必要となる．複合名詞解析の精度向上についての最近の研究では，名詞の共起情報のみでなく名詞間

表 4 問合せ (2) の場合の IDF 法による検索結果
Table 4 Search result using IDF method for the query(2)

順位	判定	文書タイトル
5	誤	「ネットワークトラフィック可聴化システムを遠隔地から利用するためのユーザインタフェース -stetho システムへの電話や WWW の応用-」
5	正	WWW におけるユーザ主導型転送データ量制御の一手法
5	正	WWW 感性検索システム KiSS におけるユーザモデルの構築

表 5 問合せ (2) の場合の本手法による検索結果
Table 5 Search result using our method for the query(2)

順位	判定	文書タイトル
4	正	WWW におけるユーザ主導型転送データ量制御の一手法
4	正	WWW 感性検索システム KiSS におけるユーザモデルの構築
26	誤	「ネットワークトラフィック可聴化システムを遠隔地から利用するためのユーザインタフェース -stetho システムへの電話や WWW の応用-」

表 6 問合せ (3) の場合の本手法による検索結果
Table 6 Search result using our method for the query(3)

順位	判定	文書タイトル
13	誤	組織間網環境におけるアクセス制御方式
16	誤	通信網ワイドの交換サービスを実現するためのプログラム制御方式
22	正	ATM 用確認型データ転送プロトコル SSCOP の実装と評価
22	正	マルチキャストサーバ方式による ATM 上の IP マルチキャスト通信

の距離やシソーラスを用いた解決方が示唆されている[19]．ただし，自然言語処理は，広範な応用を考えてなるべく負荷の小さい手法で行う必要があると考えているので，このような手法の適用には十分な検討が必要である．

一方，後者の問題は，個々に抽出した係受け関係を使ってどのように文書を特徴付けるかという問題に帰着する．本手法では係受けの一致度とそれを用いた問合せと文書との一致度の定義式にさまざまなパラメタを設定している．これらを最適化することによって，係受け関係を情報検索に対して有効に利用する方法を発見することが可能であると考えられる．また，このようなパラメタの最適化によって得られる検索精度と係受け解析の精度との関係を調べることも必要であり，そのためのテストセットの作成も必要である．

今回のテストコレクションは比較的小規模であった．本手法が，他のテストコレクションにおいても有効かどうかの実験を継続して行う必要がある．テストコレクションとしては，学術文献を対象とした NTCIR1[20] や，新聞記事を対象にした IREX[21] などがあり，これらのコレクションによる検索実験を通してシステムの有用性を評価し，改良を重ねていくことが今後の課題となる．

謝辞 本研究の端緒となった研究に参加した，細貝英篤氏（現，富士通），池田和幸氏（現，富士通）から多大な貢献があったことを記し，謝意を表します．

この研究は日本学術振興会における「高度分散情報資源活用のためのユービキタスシステムについての研究」プロジェクトの支援を受けています．

参 考 文 献

1) Donna Harman: “Ranking Algorithms” in *Information Retrieval*, Chapter 14. Prentice Hall, 1992.
2) Ulrich Pfeifer, Norbert Fuhr, Tung Huynh: “Searching structured documents with the enhanced retrieval functionality of freeWAIS-sf and SFgate”, *Computer Networks and ISDN Systems*, Vol.27, No.6, pp.1027–1036, 1995.
3) Alan F. Smeaton: “Progress in the application of natural language processing to information retrieval tasks”, *The Computer Journal*, Vol.35, No.3, 1992.
4) 那須川哲哉:「文や句による日本語テキストの検索」, 第 53 回情報処理学会全国大会, 5T-1, 1996.
5) C.Berrut: “Indexing Medical Reports: the RIME Approach”, *Information Processing and Management*, Vol.26, No.1, pp.93–110, 1990.
6) Gardin, J.C. Syntol. New Brunswick New Jersey: The Rutgers University Press; 1965.
7) Salton, G. Automatic phrase matching. In: Hays, D.G., editor. Readings in automatic language processing. New York: American Elsevier

Publishing Company Inc.; 1966: pp.169–188.
8) Lu, X.: "Document Retrieval: A Structural Approach", Information Processing & Management, 26(2), pp.209–218, 1990.
9) Sung H. M., Christopher, K., Ming, L.: "Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System", *Proc. of 2nd International Conference on Conceptual Structures*, Lecture Notes in Artificial Intelligence 835, Springer-Verlag, PP.69–83, 1994.
10) 池田和幸, 安達淳: 「表題の意味構造を考慮した文献検索手法」, 第 53 回情報処理学会全国大会, 2T-3, 1996.
11) 池田和幸, 安達淳: 「単語間の係受け情報を用いた文献検索手法」, 第 54 回情報処理学会全国大会, 4K-2, 1997.
12) 松村敦, 池田和幸, 高須淳宏, 安達淳: 「構造化インデクスを用いた情報検索システム」, *Proceeding of Advanced Database Symposium '97*, pp.151–158, 1997.
13) 松村敦, 高須淳宏, 安達淳: 構造化インデクスを用いた情報検索手法の評価, 電子情報通信学会技術研究報告, DE98-2, 1998, pp.7–14.
14) Atsushi Matsumura, Atsuhiro Takasu, Jun Adachi : "Information Retrieval Method using Structured Index for Japanese Text", *Proceedings of the 3rd International Workshop on Information Retrieval with Asian Languages(IRAL'98)*, 15-16 October 1998, p.109-115.
15) Ryoo, J.-H., Kim, D.W.,Myaeng, S.H.: "Extracting Lexico-Semantic Relations from Document Titles Using Sublanguage Analysis", *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, pp. 315–320, Phuket, Thailand, December 2–4, 1997.
16) 日本語形態素解析器 茶筌 (ChaSen) http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html
17) 宮崎正弘: 「係受け解析を用いた複合語の自動分割法」, 情報処理学会論文誌, Vol.25, No.6, pp.970–979, 1984.
18) 石崎雅人: 「複合名詞における語と語の関係について」, 第 37 回情報処理学会全国大会, 3C-2, 1988.
19) 小林義行, 徳永健伸, 田中穂積: 「名詞間の意味的共起情報を用いた複合名詞の解析」, 自然言語処理, Vol.3, No.1, pp.29–43, 1996.
20) 情報検索システム評価用テストコレクション構築プロジェクト http://www.rd.nacsis.ac.jp/~ntcadm/
21) IREX(Information Retrieval and Extraction Exercise) http://cs.nyu.edu/cs/projects/proteus/irex/

(平成 11 年 9 月 19 日受付)
(平成 11 年 11 月 26 日採録)

(担当編集委員 遠山 元道)

松村 敦 (正会員)

1997 年東北大学大学院理学研究科博士課程修了. 博士 (理学). 同年より日本学術振興会リサーチ・アソシエイトとして文部省学術情報センター研究開発部に所属. 情報検索, 電子図書館の研究開発に従事.

高須 淳宏 (正会員)

1989 年東京大学大学院工学系研究科博士課程修了. 工学博士. 1989 年に文部省学術情報センターに入所し, 現在同センター研究開発部助教授. データベースシステム, 電子図書館システム, 機械学習の研究を行っている. 電子情報通信学会, 人工知能学会, ACM 各会員.

安達 淳 (正会員)

1981 年東京大学大学院工学系研究科博士課程修了. 工学博士. 東京大学大型計算機センター助手等を経て, 現在文部省学術情報センター研究開発部教授. オンライン情報システム, 分散処理システム, 情報検索, 電子図書館システム等の開発研究に従事. 電子情報通信学会, IEEE, ACM 各会員.