特許情報検索のための日本語質問文解析†

佐藤正光†† 斎藤裕美†† 菊地紀芳††

通常の日本語文章による検索質問が可能である特許情報検索システムを開発した．本論文では，そのシステム中で用いられている質問文解析方式について述べる．本方式では，特許検索での質問文の構造的特徴を利用して，質問文中の各文節間の係り構造を認定し，その構造を用いて検索論理式を作成している．また，文章解析技術とシソーラスの利用によって，多様な質問文表現を許している．本方式をとることにより，検索者は，質問文の形式や用語表現にとらわれることなく，自由に検索質問を行うことができる．

1. まえがき

社会のさまざまな分野で情報検索システムが使われている．そのなかで，特許情報検索システムについては，そのニーズも大きく，公的機関および民間企業においてさまざまな形のシステムが作られている．

しかし，従来の情報検索システムにおいては，次のような欠点があった．

a．検索質問は式，またはきわめて制限された文（式に近い文）で行っていた．

b．キーワード付けは人手によっていた．

c．キーワードの下位語展開を自動的にしていないことが多い．

d．情報データに追加するキーワードが不統一であった（キーワード付け作業者の語いの不統一，登録者と検索者の間の語いの不統一）．

e．キーワード付けにその分野の専門家が必要であった．

また，計算機システムを計算機の専門家以外の人が使用することが多くなってきている現在，次のア～ウのような要求が生じている．われわれは，これらを満足するためには，通常の日本語文で検索質問を行えるような手段を提供する必要があると考えている．

ア．質問入力の際，難解な数学的論理式を用いずに入力できる．

イ．質問入力が日本語で行える．

ウ．語句や記述形式に制約されずに文章で質問の内容を表現できる．

そして，上述a～eの欠点，問題点を解決するべく日本語文章処理技術を用いた特許情報検索システムを開発した．このシステムは，日本語文章処理の応用システムの有用性，将来性を示したものであると考えている．

また，そのなかで，特許情報登録のためのキーワード抽出および検索質問文解析のためのキーワード抽出に共用できる文章解析プログラムを開発した．

この特許情報検索システムにおける質問文解析方式（質問文を文章解析し，検索論理式を組み立てるまでの方式）は，次の点にポイントを置いて開発した．

a．なるべく多様な日本語の質問文表現を許す．

b．特許検索の性格上，一つの用語に関連する多くの特許が検索されるようにする．

本論文では，質問文解析上の問題点，本特許情報検索システムで開発した解析方式，および実験結果について述べる．特許情報検索システムの概要については文献1）を参照されたい．

2. 特許情報検索と質問文

一般に情報検索システムにおける検索質問の形式は，次の(1)～(3)にまとめられる．

(1) 検索者が未知の数値，名称などを直接尋ねる．百科事典的性格をもつ情報検索システムに多く見られる．

例 ・東京の人口は何人ですか．

(2) 検索者が題名，日付，人名（著者名など）等のあらかじめ分類されている項目を用いて目的の事物を尋ねる．いわゆる文献検索システムがこの形である．

例 ・1975～78年発行の，数学の文献でリー群の表現に関するもの．

(3) 検索者が知っている特徴的な事物を用いて，類似の事物，同一の事物を尋ねる．あらかじめあ

† Japanese Query Sentence Analysis for Patent Information Retrieval by MASAMITSU SATOU, HIROYOSHI SAITOU and KIYOSHI KIKUCHI (Research & Development Center, Toshiba Corp.).

†† (株)東芝総合研究所

る物質の組成，製法を知っていて，それに類似な物，同一な物を知りたい場合の検索である．

例 ・白金層の上にシリコン蒸着を行った基板をもつ○○○のための回路

特許情報検索においては，(2)に類する発明者，出願日などの書誌的事項による検索質問と，(3)に類する発明内容に関する検索質問とがある．

われわれのシステムでは，前者の質問は書誌事項検索として表形式による質問入力方式をとり，後者の質問は日本語の文章によって入力する方式をとった．

(3)の形の質問文は，事物の構成や組成，製法を叙述した形の文となっている．つまり，ある事物を示す語に対し，それを叙述，修飾する語が連なって一つの文を形成しているのである．それゆえ，情報検索システムとしては質問文内の修飾，非修飾の関係，すなわち各用語間，文節間の関係を明確にして検索論理式を立てて検索することがポイントとなる．

そこで，われわれは，特許情報検索の質問文のこの特徴を考慮し，次のような条件下で検索質問を処理することにした．

① 質問は通常の日本語文章で行う．

② 質問の形式は，類似事物，同一事物の有無を尋ねる形とし，上述(1)のような質問は取り扱わない．

③ 質問文の構造は次の形に制限する．

a．文は，

修飾部＋名詞 の形とする．

b．修飾部は，名詞を説明するものであり，

{形容詞句 / 動詞句 / 名詞句} ＋接続語

の単位が連なったものである．

c．{形容詞句 / 動詞句 / 名詞句} とは，

・性質，状態，様相を記述する形容詞文節，形容動詞文節

・性質，状態，様相を記述する動詞文節

・製造方法を説明する動詞文節

・構成要素を記述する名詞文節

および，これらにさらにそれらを修飾する語や句が連なったものである．

d．接続語とは，

接続詞，接続助詞，読点類（、，；）およびこれらに準ずる語（「～とともに」など）である．

④ 検索対象は特許（公告，公開）とし，その特許番号，書誌的事項，請求範囲文，図面を応答出力する．

⑤ 質問文を解析して得られる検索論理式は，質問文から得られたキーワードを AND (*), OR (+) で結合したものとする．

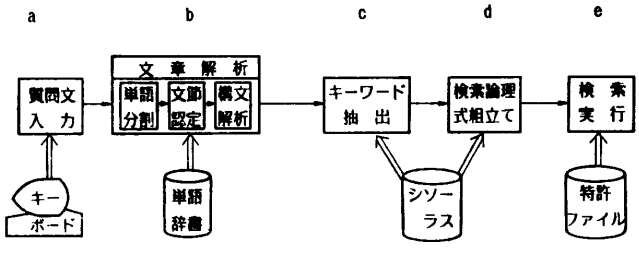

図 1 質問文入力から検索実行までの過程
Fig. 1 Process from query sentence input to retrieval.

3. 質問文の解析

われわれのシステムにおける質問文入力から検索実行までの過程を図1に示す．

図1の「文章解析」，「キーワード抽出」の過程は，質問文内の各単語，文節間の関係を明確にして，どの文節，どの単語が質問文中重要な因子であるかを把握するものである．

われわれは，特許情報検索における質問文の構造を検討し，文節間の係り受け関係を明確にすることが重要であると考え，このような方式をとった．

本章では，図1の「文章解析」，「キーワード抽出」，「検索論理式組立て」の各過程の処理方式について述べる．

3.1 質問文解析の課題

日本語質問文解析上の課題としては，次の①，②がある．

① 質問文より抽出したキーワードが妥当であること．つまり，同一の検索意図に対しての構文上の相違の吸収と，同一の検索意図に対しての用語上の相違の吸収を実現する．

② AND, OR 等の演算子の結合が妥当であること．

われわれは，これらを次のように解決した．

① 高精度な構文解析アルゴリズムによって係り受け関係を認定し，語順や表現形式の相違を吸収し，より適切なキーワードを抽出している．また，シソーラスをもち，用語の上位下位概念関係，同義関係を明らかにしている．これを用いて

同一検索意図に対する違った用語表現を吸収し，キーワードの統一化を図っている．

また，質問文より抽出したキーワードに文構造上から見たレベル付けをしている．これは，文章構造の文節間の接続関係を反映したものであり，これによって質問文の構造が検索論理式により正しく反映されることになる．

② 文節が並列しているのか，修飾関係にあるのかを正しく認定することがキーワード抽出およびAND, ORの結合の仕方に大きな影響を及ぼす．そこで，文節間の係り関係を正しく判定すること，接続語（接続詞，接続助詞を含む）の接続関係を正しく判定することが必要である．

文節間の係り受け関係の判定は①の文章解析技術によって行うことができ，接続関係の判定は，各単語に意味情報を記述し，使用範囲を明確にすることによって対処した．

3.2 文章解析とキーワード抽出

3.2.1 文章解析とキーワード抽出の手順

本節では，文章解析からキーワード抽出までの手順について簡単に述べる．詳細は文献2）を参照されたい．

まず，質問文中の文字列から，単語分割，文節認定の処理によって文節を認定する．

次に，構文解析によって，質問文中の大きな並列関係，修飾関係にある単位（おおむね形容詞句，動詞句，名詞句にまとめられる単位で，ここではこれを段落と呼ぶ）を求め，各段落内の文節の係り受け関係を求める．段落を求める際には，質問文の構造的特徴を利用している．

われわれのシステムでは，特許検索を対象としているので，質問文の構造的特徴として次の点を利用した．

a．文の終りが体言の場合が多い．

b．名詞並列が多い．

c．連用中止法が多い．

また，「持つ」，「有する」，「含む」，「備える」などの語は大きな単位の並列関係をまとめるための基本語として取り扱うこととし，分割の手がかりとした．

段落内の文節の係り受け関係の解析には，名詞の意味的分類と各動詞の要求格を記述した単語辞書を用いている．

構文解析の後，キーワード抽出を行う．

キーワード抽出には，係り受け結合法[2]を用いている．この方式では，シソーラスを用い，文節の結合順序を入れかえたりして同一概念に対する多様な表現に対処しながら，なるべく多くの単語から構成される複合語（つまり，シソーラスの構造から見ると，下位概念に属し，他の特許との区別がつけやすい語）を抽出している．シソーラスは，この方式においては，用語の多様な表現の吸収と，キーワードの統一性をはかることに用いられている．

**********　質問文　**********

気相成長層と、低比抵抗の埋込層とをもち、アンチモンまたはヒ素の不純物が拡散されているＩＣ。

*****　キーワード抽出結果　*****

『気相成長層→エピタキシャル成長層』と、『低比抵抗の埋込層』とを　：
もち、　：
「アンチモン」または「ヒ素→砒素」の『不純物が拡散』されている　：
【ＩＣ→集積回路】。　：
：　：　：
(2)　(1)　(0)

注）：および(2)，(1)，(0)は段落のレベルを示す．
→は左側の語が非代表語，右側の語が代表語であることを示す．

図2 質問文とキーワード抽出結果
Fig. 2 A query sentence and the result of keyword extraction.

図2に質問文とそのキーワード抽出結果の例を示す．

図2で，【 】，『 』，「 」で囲まれ，下線が付いているのが，キーワードである．→は，左側の語が非代表語，右側の語が代表語であることを示している．

上述の方法は，特許登録のためのキーワード抽出方法とまったく同様である．これによって登録の際に付加したキーワードと検索の際に用いるキーワードの不一致が起こらないようにしている．

3.2.2 キーワードの区別

図2で，【 】，『 』，「 」はキーワードの文構造上の係りの関係を反映をしたものである．これらは，次のように区別している．

【 】：文構造上，最終的に修飾されている語に付ける．すなわち，質問文を修飾語＋名詞と考えたときの名詞に相当する語に付ける．特許の名称に相当する語である場合が多い．

『 』：各段落内において最終的に修飾される語を重点にして，係り受け結合法によって抽出されたキーワードを示す．検索したい事物の特徴を記述した語である場合が多い．

表 1 接続語のカテゴリの例
Table 1 An example of category of connectives.

カテゴリ	接続語
OR (+)	あるいは，または，もしくは，～のいずれか
AND (*)	かつ，および，及び，～とともに

********** 論理式 **********
(『エピタキシャル成長層』*『低比抵抗埋込層』*(「アンチモン」+「砒素」)*『不純物拡散』)*【集積回路】

図 3 図2の質問文から作られる検索論理式
Fig. 3 The logical expression for retrieval made from the query sentence shown in Fig. 2.

「 」:『 』で示すキーワードを抽出後，同じ段落内の残りの語からさらに同じ手順で抽出されるキーワードを示す．『 』で示されたキーワードを補足するものである．

したがって，キーワードに付けられたカッコを調べることによって，キーワード間の従属関係を知ることができる．

3.3 検索論理式の組立て

質問文からキーワード抽出後，検索論理式への組み立てを行う．構文解析結果，キーワード抽出結果とキーワードを含む文節を接続している接続語についての意味カテゴリ(**表1**参照)とを用いて，AND (*)，OR (+) の論理記号と各キーワードを結んで検索論理式を組み立てる．

たとえば，図2の例からは，**図3**の検索論理式が作られる．

3.3.1 項にそのアルゴリズムを示す．

3.3.1 検索論理式の組立てアルゴリズム

検索論理式を組み立てるときには，質問文解析からキーワード抽出までで得られた情報のうち，構文解析結果，キーワード抽出結果を利用する．

われわれのシステムでは，以下のような手順で論理演算記号 (*, +) をつけてキーワードを結合する．

① AND (*)

a．暗黙的な AND (*で示す)

原則として，とくに以下の規定 (bおよび②) に該当しないすべてのキーワード (それを含む文節や段落) は，この記号によって結合される．

b．決定的 AND (※で示す)

かつ，および (及び)，とともに，という接続語によって文節，段落が接続されているときは，この記号によってキーワードを結合する．

※は最終的な検索論理式では*に置換する．

② OR (+) (決定的 OR, +で示す)

または (又は)，あるいは，という接続語によって文節，段落が接続されているときは，この記号によって，キーワードを結合する．上記の接続語は，それらの語以前の上記接続語に係っている文節，段落，キーワードを結合している*を+に変える働きをもつ．

次に手順を述べる．

(1) 各段落 (または，段落内の同一レベルのキーワード群) を( )でくくる (以下，この単位をブロックと呼ぶ)．

(2) すべてのキーワードを*で結合する．

たとえば，

「金，銀，銅を含み，…」

というときには，

(『金』*『銀』*『銅』*…)

の形に結合される．

(3) 構文解析結果を参照しながら，※と判断されるところは，※に置換する．

たとえば，

「金を含む層を作り，銅およびアルミニウムの電極を…」

というときには，

(『金』)*(『銅電極』※『アルミニウム電極』…

の形となる．

(4) ブロックの文節の終りに OR と判定される語があるときは，そのブロック内のその語よりも前の*はすべて+に置換する．ただし，※のある個所は( ) でくくって※を残す．

たとえば，

「金，銀，銅のいずれかを含み…鉄およびコバルトの層またはアルミニウムの層…」

のときには，

(『金』+『銀』+『銅』)*…((『鉄』※『コバルト』)+『アルミニウム』…

となる．

(5) 各ブロック間についても，( )でくくられた部分全体を一つのキーワードと考え，(1)～(4)の処理を行う．

(6) でき上がった式を走査して，※は*に変える．

3.3.2 多様な質問文への対処

われわれは，3.3.1 項で述べた方法によって検索論理式を組み立てている．そして質問文内の論理関係が十分検索論理式に反映され，しかも多様な質問文の

表現を許すために次の方法をとっている.

ア. 質問文の文節の係り受け構造を反映したキーワードの区別.

イ. 係り受け関係を認定し，語と語の間の結合関係を認識することにより，同一の検索意図に対し多様な質問文の表現を許す.

ウ. 質問文中の用語の意味的相違，同義関係をシソーラスから求め，検索意図に合ったキーワード抽出を行う.

われわれの方式は，上記の方法により，構文形式，用語表現の相違は吸収しているが，大小関係や数値表現の相違については対処していない. また，多様な質問文といっても，大小関係や数値自体はキーワードとしていないのでこれらについての質問文は取り扱っていない.

この2点については，今後の課題と考えている.

3.4 シソーラスと下位語展開

われわれのシステムにおいては，多様な質問文表現に対処するためとキーワードの統一をはかるために，シソーラスを使用している.

このシソーラスは，特許登録時のキーワード抽出，質問文解析時のキーワード抽出，どちらの際にも使用される. シソーラスの構造の概略を図4に示す.

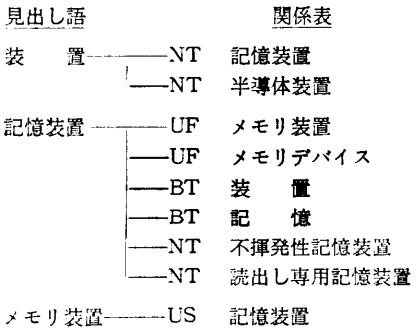

図4 シソーラスの構造の概略
Fig. 4 An outline of the structure of the thesaurus of the system.

＊＊＊＊＊＊＊＊＊ 論理式 ＊＊＊＊＊＊＊＊＊
『電界効果トランジスタ』＊【メモリ装置】

＊＊＊＊＊＊＊＊＊ 下位語展開 ＊＊＊＊＊＊＊＊＊
（『電界効果トランジスタ』＋『接合電界効果トランジスタ』＋『絶縁ゲート電界効果トランジスタ』＋『相補電界効果トランジスタ』＋『MIS電界効果トランジスタ』＋『逆導電形MIS電界効果トランジスタ』＋『MIS電界効果トランジスタ領域』＋『浮遊ゲートMIS電界効果トランジスタ』＋『MOS電界効果トランジスタ』）＊（【記憶装置】＋【半導体記憶装置】＋【不揮発性記憶装置】＋……）

図5 検索論理式とその下位語展開の例
Fig. 5 An example of a logical expression for retrieval and the keyword development with narrower terms.

オペレータが入力したキーワードや質問文中の用語で，代表語でないものはこのシソーラスを見て代表語に変換される.

シソーラスを用いることによって，質問文に使われる用語の相違を吸収し，質問文の多様な表現を許している. このほか，より広い検索を行い，検索もれを防ぐための下位語展開にもこのシソーラスが使われる.

下位語展開は，検索もれを防ぐための手段であり，われわれのシステムでは自動的に行っている.

展開のレベルは，今回は1段下位までとしたが，論理的にはオペレータの指示により n 段下位まで展開可能である.

図5に下位語展開の例を示す.

下位語展開の効果としては，検索もれの減少のほか，検索質問のなかで，求める特許についていると予想されるキーワードを逐一指定しなくてもよいことが上げられる. 実験によれば，キーワードにもよるが，下位語展開をした場合の検索結果件数は，しなかった場合の約2倍である.

下位語展開が有効であった質問文は，

ア. 方法，方式に関するもの

イ. 生成物，構成に関する質問で，該当する生成物や構成が多種類のもの

のように広い範囲を包含する形のものが多い. 生成物や構成に関する質問で，該当する生成物や構成の種類が少ないもののときには，関係のない特許が検索結果に含まれることがしばしば見られる.

これは，下位語展開をするキーワードについて，それを含む文節と他の文節との関係を反映せずに単純にそのキーワードを下位語展開しているので，関係のない下位語までが展開結果に含まれるためと考えられる. したがって，検索結果の質を向上させるためには，下位語展開の際にも文節間の関係を生かした展開を行う必要がある.

たとえば，「金を拡散させる工程を含む半導体装置の製造方法」という質問文では，キーワード『金拡散』を下位語展開するとき，工程に注目し，工程に無関係な下位語を省くといった措置をとるのである.

4. 解析例と考察

本章では，質問文の解析例を示し，考察を加える（図6参照）.

① ＊＊＊＊＊＊＊＊ 質問文 ＊＊＊＊＊＊＊＊
無孔質アルミナを被膜し、断線を防止した半導体装置の製造方法。
＊＊＊＊＊＊＊＊ 論理式 ＊＊＊＊＊＊＊＊
（（『無孔質アルミナ被覆』＋『断線防止』）＊【半導体装置製造法】

② ＊＊＊＊＊＊＊＊ 質問文 ＊＊＊＊＊＊＊＊
シリコンカーバイトまたはガラス層を有し、その上に受動素子と能動素子がともに形成された半導体装置。
＊＊＊＊＊＊＊＊ 論理式 ＊＊＊＊＊＊＊＊
（（『シリコン炭化物』＋『ガラス層』）＊『受動素子』＊『能動素子』）＊【半導体装置】

③ ＊＊＊＊＊＊＊＊ 質問文 ＊＊＊＊＊＊＊＊
埋没気相成長層を有し、PNP、NPNTr を備える半導体装置。
＊＊＊＊＊＊＊＊ 論理式 ＊＊＊＊＊＊＊＊
（『埋込み気相成長層』＊『PNPトランジスタ』＊『NPNトランジスタ』）＊【半導体装置】

④ ＊＊＊＊＊＊＊＊ 質問文 ＊＊＊＊＊＊＊＊
埋込み層とともに、N型基板、P型基板、半導体薄板のいずれかを有する半導体装置。
＊＊＊＊＊＊＊＊ 論理式 ＊＊＊＊＊＊＊＊
（『埋込み層』＊（『N型基板』＋『P型基板』＋『半導体薄板』））＊【半導体装置】

⑤ ＊＊＊＊＊＊＊＊ 質問文 ＊＊＊＊＊＊＊＊
凹みを形成し、この凹み内に半導体領域または半導体層または気相成長層を設ける工程を含む半導体装置の製法。
＊＊＊＊＊＊＊＊ 論理式 ＊＊＊＊＊＊＊＊
（『凹み』＊（『半導体領域』＋『半導体層』＋『気相成長層』））＊【半導体装置製造法】

図 6 解析例
Fig. 6 Examples of query sentence analysis.

①は，基本的な解析結果例である．

①において，“無孔質アルミナ被覆”という語がシソーラスに登録されておらず，“無孔質”と“アルミナ被覆”がシソーラス上に登録されていると，検索論理式は，（(「無孔質」＊『アルミナ被覆』)＊『断線防止』)＊【半導体装置製造法】となる．また，質問文が，“…アルミナ被膜を形成し…”，“…アルミナ被覆を設ける工程を含み…”，“…Al 被膜をもち…”などの表現をとっていても，検索論理式は上記と同様になる．

②は，質問文に並列表現がされている場合の例である．

“受動素子と能動素子がともに…”の部分は“受動，能動素子が…”，“受動素子および能動素子とが…”または“受動素子，能動素子を形成した”などと表現しても同じ検索論理式が作られる．

“形成する(される)”，“有する”などは“その上に”などと同様，文章解析の際の手がかりとして取り扱い，検索論理式作成の際にはその語のもつ意味関係を解釈していない．

④では，「いずれかを」という接続語によりその前のキーワードをすべて AND(＊) 結合ではなく，OR(＋)結合としている．

以上のように解析例からみると，用語や接続語が多い例も正しく検索論理式を組み立てている．用語の違いも含めた質問文の表現の相違も許されている．このように，当初の目標である「通常の日本語文を用いて検索質問を行う」ことは解析例に示した程度であれば十分達成できたと考えている．

ところで，いま少し深く検索質問について考えてみると，次のような問題がある．

「…金または銀の金属が拡散されている…IC」という特許請求範囲文の特許を登録し，キーワードとして「金」，「銀」，『拡散』，【IC】が付けられたとする．そして，「金または銀の不純物が拡散されている IC」という質問をしたとすると，現在のわれわれの方式では，検索論理式は，(「金」+「銀」)＊『不純物拡散』＊【IC】となる．したがって，先の特許は検索されない．

しかし，質問文をよく考えてみると，検索者は，「拡散工程において，あるいは IC の組成上，金，銀は不純物として位置づけられる」ことを知っていると考えられる．すると，先の特許は，“不純物”という用語がなくても検索されなければならないことになる．

このような問題に対処するには，IC と金・銀の組成上の関係や IC の製作上の知識を検索システムがもち，これに応じて自動的に検索論理式を修正して検索したり，検索者に対し検索論理式が (「金」＊「銀」)＊『拡散』＊【IC】と等価であることを教えて指示を仰ぐといった方法をとらなければならない．

5. 結 び

日本語文による質問文解析について述べた．

われわれの方式は，文章解析技術を用いて構文解析，キーワード抽出を行った後，文節の接続関係から検索論理式を組み立てるというものである．解析例にも示したとおり，十分実用に耐えるものと考えている．

われわれの方式によれば，用語や形式にとらわれることなく，質問文を作成することができる．すなわち

① 質問文中の文節の係り構造を認定し，自動的に論理関係を導出しているので，質問文を表現する際に検索論理式を記述する必要がない．
② 質問文を作る際，正確に用語を記述しなくても正しい用語に変換されるので，使用する用語，その語順，表現形式にとらわれなくてよい．
③ 下位語展開を自動的に行っているので，詳細に

用語を記述する必要がない.

今後の課題としては,

a. 特許の構成手順, 特許中の数値 (たとえば, 温度, 圧力, 元素の含有率など) に特許性があって, これに重点を置いて検索したいときにどのようにすればよいか,

b. 文章解析, シソーラス参照では把握できない用語の概念上の相違, 類似点に中心を置いて検索をしたいときにはどうすればよいか,

c. 否定語の処理はどのようにするのがよいか,

があげられる.

また, 特許中の用語は年月がたつにつれてその呼称が変わっていくことが多いので, これらの管理を正しく行うことが重要である.

なお, 本研究は通産省工業技術院の大型プロジェクト "パターン情報処理システムの研究" の一環として行ったものである.

参 考 文 献

1) 佐藤, 菊地, 野寄, 斎藤: 日本語による特許情報検索システム, 情報管理, Vol. 24, No. 4, pp. 343-352 (1981).

2) 斎藤, 野寄: 特許請求範囲文の文章解析によるキーワード抽出, 電子通信学会論文誌, Vol. J 65-D, No. 10, pp. 1195-1202 (1982).

3) 金平: JAPATIC のオンライン特許情報検索システム (PATOLIS) について, 情報管理, Vol. 21, No. 9, pp. 681-684 (1978).

4) 長尾, 辻井, 田中: 意味および文脈情報を用いた日本語文の解析, 情報処理, Vol. 17, No. 1, pp. 10-28 (1975).

(昭和 58 年 1 月 27 日受付)
(昭和 58 年 10 月 11 日採録)