Library and Information Science No. 31 1993

段落を対象とした日本語全文データベースの検索

Full-text Database Retrieval Using Paragraphs: In the Case of Japanese Technical Document Database

野　末　道　子
Michiko Nozue

Résumé

In these days the online full-text databases are increasing, but these full-text databases are difficult to retrieve, because recall is higher than bibliographic databases, and precision is so lower. There are cases where we don't always read whole paper, but use one or a few part of an article. So this paper presents an approach to retrieve the relevant parts of a document by using paragraphs of individual documents. Sample documents are 49 articles in Japanese about information retrieval and natural language processing studies. The retrieval technique used in this retrieve experiment is the vector space model. As a result, the higher precision and recall were shown by using the words in chapter titles or section headings to retrieve the relevant paragraphs.

野末道子：鉄道総合技術研究所，池田(知識データベース)研究室，国分寺市光町 2-8-38
Michiko Nozue: Railway Technical Research Institute, Knowledge & database engineering laboratory, 2-8-38, Hikari-cho, Kokubunji-shi, Tokyo.
1994 年 2 月 5 日受付

I. 全文データベースの現状と問題点

これまで情報検索分野では，主として書誌データベースを対象とした検索手法が研究されてきた。しかし，全文データベースの発展にともなって，全文データベースの検索手法の研究へと焦点が移りつつある。

全文データベースは書誌データベースとは二つの点で大きく異なっている。一つは全文データベースでは，提供される対象が全文となるという点である。書誌データベースでは，主題探索の際の対象は，標題，抄録，それに付与されたキーワード群，あるいは引用文献などである。もう一つの相違点は，検索と同時に全文が入手できるという点である。一方の書誌データベースでは，検索した後に別の手続きによって求める文献を入手する必要がある。

しかしながら，こうした二つの特色が区分されないまま「全文データベース」という表現が用いられており，次のような種類のデータベースが全文データベースと見なされることが多い。

①文章中の語及び，図表，写真なども検索が可能なデータベース
②図表などのタイトル及び文章中の語でのみ，検索が可能なデータベース
③全文が蓄積してあるが，書誌，抄録中の語でのみ検索可能なデータベース

①は実例は乏しく，②が一般的であり，現在ほとんどのオンラインデータベースは，図，表，写真などの画像へのアクセスポイントの付与，及び出力の問題は未解決となっている。②の例では，オンライン情報検索サービス機関がこれまでの書誌ファイルと共に提供する *Harvard business review* などの論文や，国内外新聞記事などの全文のファイルと，CD-ROM の形態で出版されたものが挙げられる。③の例としては，「Business Periodicals On Disc」や「ADONIS」のサービスのように書誌データベースとイメージ形態の全文とを同時に提供する CD-ROM もある。

しかし，このような全文データベースの増加とハードウェア面の進歩が著しい一方で，全文データベースの検索手法の研究はまだそれほど進んでいない。今後全文検索システムを構築していく上では，全文データベースの役割，及び特徴を改めて考えていく必要がある。

本論文では，一次情報の提供ができるという全文データベースの特徴に焦点を当て，ここから考えられる検索システムの可能性について検討を行った。文献は複数の主題から構成されているという点を考え，検索結果として提供されるのは，1 文献単位ではなく，主題を表現している部分でよいという立場をとった。この部分の検索のため，部分を表現するよりよい索引方法を検討し検索実験を行った。

本論文では，II 章において索引方法及び日本語特有の問題点について述べる。そして，III 章でこれまで行われてきた全文データベースの検索実験研究，及びこの論文の仮定となる部分を対象とした検索の意義について考え，IV 章以降で実験とその結果，考察を行うこととする。

II. 全文データベース検索システムと索引手法

全文データベースでは書誌データベースに比べ，より多くのアクセスポイントが提供されるために再現率が向上する，ということが一般的な見解となっている。しかしこれにともなって，全文データベースでは多くの不要な文献も検索される出力過多の問題も指摘されてきた。出力過多が生じると論理積を用いて検索結果を限定することから，かえって実際の再現率は低下すると Blair らは述べている。[1)2)] これは，本文の主題とは直接関係のない概念や，本文中で特に重視されているわけではない概念を表しているような語を含め，すべての語をキーワードとしているという索引方式の問題点として考えられる。この問題は抄録についても当てはまるが，特に全文データベースの場合において顕著である。そのため，キーワードを認識するような抽出索引方式，もしくは自然言語の表記上の揺れを修正するための付与索引方式が必要となると考えられる。

英文などの区切りのある言語においては，語の切り離

しの問題は生じない。しかし検索速度の向上のためインバーテッドファイルを用意する場合には，キーワード抽出のための辞書は必要であり，複合語をどのように取り入れるかということも問題となる。日本語文のような膠着言語においてキーワードを抽出する場合には，ここで用いる辞書の質，文章解析能力が必要となり，辞書の質の維持には多大な労力が要される。当然のことながら，キーワードのインバーテッドファイルを用いずそのまま全文サーチをする場合には辞書は不要であり，いかなる複合語を使って検索することも可能である。

以下では現状の全文データベースはどのような索引方法を採用しているかについて概観する。全文データベースの索引方式は，現在のところほとんどが，出現した語をそのままフリーキーワードとして採用するものである。そこで，抄録中の語を対象として機械的に語を抽出しているものについても，全文を対象として応用できるものは含めて以下で述べる。

A. キーワードの抽出と重み付け

文献に与えるキーワードを決定する方法としては，付与索引方式と抽出索引方式がある。付与索引方式は，対象文献中の主題分析を行い，個々の主題概念を的確に表現するキーワードを，文献中で使用されている語とは別に付与する方式である。この付与索引方式で与えられるキーワードは，シソーラスなどの用語集や，分類表，件名標目表から選択されることが多い。このようにして与えられるキーワードは語彙統制がなされている。しかし付与索引方式には，知的判断が必要であることから，自動索引向きの方法とはいえない。一方，抽出索引方式は，基本的に文献中に出現した語をそのままキーワードとして使用する方式である。この方式では，索引作成者の経験や専門的知識を要するものではないことから，自動化の研究が活発に行われている。[3)]

この中で，確率・統計的手法を用いてキーワードを抽出する方法，文章を構文解析し適切なキーワードを抽出する方法が，抽出索引研究の主流となっている。また近年では分野の専門知識を用いたエキスパートシステムによりキーワードとして適切なものを推論する研究もなされている。前者の統計的，確率的検索技術の研究では，自動索引，自動抄録，自動分類，自動探索の一連の流れについて実験を行っている，Salton の SMART 検索システムが代表的である。

SMART 検索システムにおいて，文献は重み付けがなされたキーワードのベクトルとして表され，多次元空間に配置されるものと考えられている。それぞれの語の重みが正である場合には，そのキーワードが文献に実際に付与されており，重みが 0 である場合には付与されていないことを意味している。また個々の質問も同様に，質問に対して用いられる語のベクトルとして表現される。検索は質問と文献の類似度に基づいて行うが，これは各々のキーワードを座標軸としコサイン関数を用いることにより計算する。これは，文献と質問をキーワードベクトルで表現し，文献間，もしくは文献と質問の間の角度のコサインを計るものである。

SMART 検索システムにおける自動索引システムでは，シソーラス，語の階層配置，意味解析，構文解析を行う機能を含んでいる。そして対象となる文章から切り出される語から，ストップワードの除去，及び以下に述べる語の重要度識別モデルを利用した重みの付与を行っている。

Salton が語の重み付けのために用いている重要度識別モデルは，(1) 文献中の語の頻度と，(2) データベース中のその語を持つ文献の出現頻度，及び (3) 文献の長さの三つの要素を組み合わせたものである。このアルゴリズムの妥当性については，6 種類のデーベースを使用して SMART システムで実験が行われ，これら三つの要素をすべて取り入れたモデルが最も有効であることが示されている。[4)5)]

B. 日本語文検索システムと自動索引

日本語ワードプロセッサを含む漢字入出力機器の普及と，日本語情報処理技術の進歩により，漢字や平仮名を含む日本語の情報，文書情報を対象にした情報検索システムについても，関心が高まってきている。全文データベースからのキーワードの自動抽出のレベルを高めるためには，自然言語処理の成果の利用が必要である。本節では，日本語における自然言語処理の手法および，この成果を利用した検索システムの例を検討する。

日本語文献の処理には，英語文献を処理する場合とは異なった，日本語固有の特徴がある。長尾らが顕著なものとしてあげているところでは，

①文が語毎に区切られていない
②使用される文字が漢字，平仮名，片仮名，ローマ字など数千種にのぼる
③漢字が表意文字であるため，熟語，複合語を作る造語性が高い

④外来語が片仮名で表され，表記の揺れを生じ易いなどがある。[6]

この他にも問題点として，文法的な文章構造が曖昧なものとなりやすい点も考えられる。キーワードを抽出する際には，これらの特徴を考慮しなければならない。現在，様々な自然言語処理の手法が開発されているが，これらの手法をどのように検索システムに取り入れていくかが日本語全文データベースの検索性能に直接関わってくる。

このように複雑な日本語文を解析し，キーワードを抽出するための前提となるのは電子化辞書である。横井は電子化辞書を以下の3点の条件により定義している。

①コンピュータが処理可能な機械可読辞書
②コンピュータが(自然)言語を処理理解するために用いられる機械可理解性辞書
③最新のコンピュータ技術や自然言語処理技術を用いなくては実現されない程の規模と精度が要求される辞書

電子化辞書を利用する目的は，コンピュータによる文書の作成，蓄積・検索，翻訳，要約，伝達などである。電子化辞書は処理される対象の文章に含まれる語彙を全て含む事が望まれる。一部の語彙だけしか含まないものであれば，実際には利用価値のないものとなってしまう。また新たな処理文章の増加に伴って，辞書の不断の保守や自動拡張のための支援システムが必要である。しかし単に語数が多ければ良いのではなく，形態的な情報や構文的な情報にも高い処理精度が要求されることとなる。これらは，通常の辞書ではそれほど注意の払われない部分であるが，電子化辞書では文章を解析するための唯一の手がかりとなる重要なものである。[7]

電子化辞書の作成に関わる近年の研究としては，一般用語だけからなる国語辞書を用いて形態素解析を行い，未登録語となった専門用語を抽出する吉村の研究や，[8] あるしきい値以上の頻度の語を基礎用語とし，文献集合における共出現をもとに，用語間の概念の階層関係を導入した知識ベースを構築して，専門用語集の作成支援を行っている小西らの研究がある。[9]

長尾らの実験では，特に辞書といったものを使わずにキーワードを抽出している。先にあげた日本語の特徴である文字種の違いという点を利用すると，おおまかな重要度の判定ができるようになる。具体的には，文中に現れる漢字，片仮名を拾い読みするだけで大意をつかむ事ができ，一方平仮名部分は重要度が低いと考えられる。もっとも，平仮名混じりの名詞などもあり，漢字，片仮名部分のみ注目したキーワード抽出では完全なものとはいえない。

長尾らは日本語文献におけるキーワードを抽出するために，文献集合をいくつかの分野に分け，それぞれの分野毎に現れるすべての語の出現頻度を求め，特定の分野のみに現れる語を重要な語と考えている。その際の重要な語を判断するために，カイ二乗分布による検定法を利用している。ここでは各々の文献中に含まれる名詞語の出現頻度を分野毎の標本値とし，帰無仮説として「その単語の出現する確率は全分野を通じて等しい」と仮定する。そしてここからカイ二乗値を求め，これが十分大きい値であれば「分野によって頻度に偏りがある」ということになる。このようなある分野にはよく現れるが他のほとんどの分野にはあまり現れない単語であれば，そのような単語を「特定分野を特色づける単語」としてキーワードとしての資格が与えられるものと考える。

この考え方に基づき，中学校理科の教科書と，科学技術文献速報：電気工学編の抄録部分を対象として名詞語を抽出し，カイ二乗値の大きい順に語を並べることで，キーワードをどの程度抽出できるかという実験を行っている。この結果から，カイ二乗値で上位にランク付けされた語は，ほぼ各分野を特徴づける重要な語であるという結果となっている。しかし，一般的な語が上位に残っていたり，複合語をどのように切り離すべきであるかという問題も生じている。[6]

絹川らは，日立製作所で開発されている汎用日本語情報検索のソフトウェアである ORION を中心とした日本語情報検索システムの構築を行い，自動索引方式の実験を行っている。索引方式として，日本語文構造解析方式と不要語除去方式を設定している。

この処理は，最初対象文が入力されると文字種が変化する時点で「文節」として分割し，その文節と自立語辞書と付属語表を照合させ，名詞，動詞，付属語の認定を行う。次に，日本語文型表を参照して文構造を認定し，当該文節の支配する動詞，当該文節を構成する名詞の意味，当該文節につく付属語から判断したロールを付与する。ここで付与されるロールは，①主体，②客体，③時，④場所，④活動，⑤その他の主題となっている。そして最後にモニター端末を見ながら，対話型で修正を行うことができるようになっている。

一方，不要語除去方式については，日本語文では英文のように語単位で分かち書きされているわけではないた

め，構文解析方式のように何らかの語への分解過程を必要とする。この実験では，日本語文で重要な語は平仮名以外で書き表されているという特性を利用して，文字列の分割を行っている。これをもとに，付属語表を用いて分かち書きされた文節から付属語を除去することにより自立語を認定する。またここから，副詞，連体詞，接続詞や名詞・動詞の中から，処理対象分野において情報的に重要ではない語を不要語テーブルとして登録し，これを用いて不要語の除去を行っている。この方式では適用分野による各辞書の登録語彙の異なりや，抽出アルゴリズムの種類については，どのような分野であっても同じものを適用できるという利点がある。[10]

木本は，キーワード抽出の精度をさらに向上させるため，従来行われてきた手法によってキーワードの候補となる語を抽出し，これにキーワードらしさの評価値を与えて順位付けし，キーワードを相対的に評価する INDEXER と呼ばれる実験プログラムを作成し評価を行っている。この判定，評価を行うため，①並立に表現された語，②連体修飾語，③強調表現された語，④シソーラスにおける上位下位関係にある語，⑤語の文章中における出現位置，⑥出現頻度といった語の特徴を抽出している。評価のために用いているのは一般新聞紙の全ての分野の記事であり，さらにこれらの分野の論理展開の特徴を利用した手法の評価も行っている。[11]

同様に，評価点数法を取り入れている例としては，中村らの研究が挙げられる。実験は書籍や社内報告書などの，それぞれ形態の異なる文書が選ばれており，書籍の場合は巻末索引，他のものについては執筆者の選定したキーワードを比較対象としている。中村らは簡単な不要語辞書を用いて不要語の削除を行った後，語の出現頻度や出現パタン等の統計的な情報と文法的な情報に基づいてキーワードの評価を行っている。

中村らの語の統計的な情報については，ほぼ Salton の SMART 検索システムにおいてとられている手法を土台としている。しかし，ここでは Salton の手法に加え，重要な主題が文書の中で何度も繰り返して取り上げられるという，文書内での語の出現特性を考慮している。この文書内での語の出現特性は，ある主題が文章中において重要である場合，主題を提示する段落，主題を展開する段落，また主題を再度提示する段落というパタンを取ることが多いというものである。この場合，重要な概念を表す語の出現は文書全体に及び，しかも詳細に展開する段落の中で繰り返し出現することになるため，出現する位置は不均衡となる。これに対し，一般的な語は広い範囲に分布したとしても，その出現位置が特定の段落に局所的に偏ることはあまりないと考えられる。こうした傾向を見るために，分散度，分布，頻度の3種のパラメータを取っている。

一方の文法的な情報と述べているのは，文の主題及び読者の注意を促す形態上の特徴を文章の解析時に抽出し，これに評価値を与えるというものである。ここで抽出される特徴としては，

①タイトルもしくはサブタイトルに含まれる場合
②特定の助詞もしくは助詞相当語に連接している場合
③箇条書き（名詞のみでなされている場合）
④括弧書きなどの記号が使用されている場合

などが挙げられている。[12]

また，検索式の中で得られたキーワードだけでは検索の意図が十分に反映されないことが多い。日本語の記述で問題とされるのは，表記の揺れが非常に多いことであり，これは，トランケーションなどの手法では解決できないものである。たとえば，「にほん」をキーワードとして検索する場合，「にっぽん」というキーワードで蓄積されているものは検索されない。外来語については，特に「デジタル」と「ディジタル」，「ファジー」と「ファジィ」などのように表記の揺れが問題となる。これについては，国語審議会では「外来語の表記」を作成しているが，表記上のよりどころとする性格のものであってその使用を強制をするものではないため問題が解決されるものではない。また，「計算機」をキーワードとして検索しても「コンピュータ」は検索されないなど，外来語と漢字が混在しているために，同義語が一層多様化している。これらの問題を解決するために，索引作業と検索補助ツールを連動させた研究が行われる必要がある。

III. 全文データベースの検索手法

A. 段落を中心とした論理構造と SGML

文章は，読者の理解を促すために，意味的まとまりに応じて何らかの形で分割が行われる。これは目に見える構造である段落（形式段落）として明示されている。

また文章の長さによって，段落分割はさらに，章，節，項などといった木構造の形態をとって表現されている。これらのいわゆる「論理構造」は，テキストを読み進める上での重要な手がかりとなっている。

論理構造は，著者の表現しようとする複数の主題を，階層構造に従って配置したものであり，主題間の関係も

また，この構造から把握することが可能である。この階層構造においては，章，節，段落，文，語の順に従って，主題がその細部へと展開していく。

しかしこれらの論理構造は，人間が目で見て認識することは容易であるが，文章を電子的に蓄積する場合，コンピュータで自動認識するための手段が必要となる。この電子的文書交換のための手段として論理構造を認識させるSGMLが設定されている。SGMLを取り入れて記述されたデータベースは，文章の論理構造を各要素別の区切り記号であるタグで囲むことで示し，論理要素の判別・抽出が，機械的にできるようにしたものである。近年，SGMLを取り入れて，文書を記述し，蓄積している例がいくつか見られる。米国OCLCの化学百科事典はその一例であり，この事典には，図表，テキスト，数式などが含まれ，SGMLを利用して記述されているので，段落，章，項目，ページや参照を自由に検索できるようになっている。[13] また学術情報センターにおいても，全文データベース提供のためにSGMLを用いた電子化が進められており，[14]『情報知識学会誌』などの国内の学会誌製作にも利用されはじめている。[15]

現在，全文データベース化の対象となっているものとしては，新聞記事や雑誌論文を挙げることができるが，これらは，SGMLが対象とする文，段落，あるいは章，節という単位で分割ができる。全文データベース検索の課題の一つは，こうした論理構造を生かした検索手法の開発であると考えられるようになってきた。

しかし，段落をはじめとする章，節などの論理構造を主題の一側面を表す意味的まとまりとして認識するには，不安定な要素もある。文が論理構造の一つの構成要素となって成り立っていることは，音声言語，文字言語の両面から異論のないところとなっている。また，一方の段落については文章研究においては多くの議論がなされ種々の説がある。

段落の定義について長尾がまとめているものによると，(1) 文章を構成する部分として区分され，主題を支える論点や材料を述べる小主題をもって統一されている文集合の切れ目，または，その文集合体の全体，(2) 文章の表現意図をよりよく伝えるために区分された，内容上・形式上のひとまとまりの部分，といった見解が一般的である。後者の見解にあるように，この段落については改行によって形式上区切られた形式段落，内容上の観点からいくつかの部分をまとめた意味段落（内容段落）の2種類がある。[16]

書き手独自の傾向や全体的な文章の長さにより，形式段落で区切られる文章の長さにはかなりの相違がある。特に最近の文章では頻繁に改行しているものが多く，意味的まとまりとして成り立たないような段落分割がなされている場合も数多く存在する。そこで，文章の主題構成を考えるときには形式段落だけでは限界があり，内容段落での区分で考える必要がある。[17][18]

しかし，こうした意味段落の認識は個々の書き手，読み手の認識構造に基づくものであり，いかなる場合においても絶対のものとすることはできない。そこで，永野らと同様，“文章の構造を解明するために手がかりとすべきものは形式段落である”[19] と考え，本研究においても，「形式段落は小主題のもとで文がまとまったものである」と考え，形式段落を用いた検索実験を行うこととする。

B. 全文データベースの利用者と利用形態

全文データベースの利用者は，特定の主題についての情報を要求する人々と考えられ，この点においては書誌データベースの検索者との重なりがある。また，書誌データベースを検索し，この検索結果をもとに全文データベースにもアクセスしようとする利用者も想定される。したがってこのデータベース利用者は，学生から，研究者，一般の利用者など幅広いものとなる。

これまでは，全文データベースのシステムの現状については多くの研究がなされてきた。しかし全文を蓄積し提供する上で，どのような利用者が，どのように全文データベースを利用するのかといった調査はあまり行われていない。また，利用者をとりまく研究環境やオフィス環境など様々な利用者調査も必要である。

この調査例として，Dillonらは，現在，研究者がどのように文献を読んでいるのかという文献利用方法の調査を行い，ここから考えられる全文データベースの設計への指針を示している。[20] この調査では，二つの観点から，心理学系統の研究者を対象に，雑誌の利用頻度や利用する主題領域，コピー状況などの簡単な個別のインタビューと，雑誌の読み方のプロトコルアナリシスを行っている。インタビューでは，提案されている利用可能な全文データベースに対しては，全員が今の雑誌の利用におきかわるものとしてではなく，これを補うものとして利用すると考えていることが明らかになっている。また，雑誌の読み方のプロトコルアナリシスでは，大部分の被験者がまず最初に目次に目を通し，関心のある論文

が見つかった場合に，その論文の開始ページを開き，標題と著者による確認を行っている。この後，抄録にはざっと目を通す程度の人の割合が高く，多くの被験者が抄録に批判的な見解を持っている。次に本文に目を通す過程では，最初の導入部と，章節の見出し，図表，結論からその論文の価値を判断するものが多い。この本文に加え，引用文献や，著者の所属機関も，論文の質を推測する上での判断材料となっている。そして，実際に全体に目を通すかどうかの判断は，内容がレレバントであるかどうかという問題が最も大きいものであるが，節に分かれていない，方法と結果の節が長い，考察が短い，論文全体が長いといった要素も関係している。これらの判断の結果，論文を読むと決定した場合，その論文をどのように読むかという読み方には二種類あると分析している。これは

①適切な情報をすばやく引き出すために，物理的順序とは別の方法で目を通す

②最初から終わりまで順序通りに詳しく読む

というものである。ここで注目されるのは，①の物理的順序とは別の方法で論文を読むという方式であるが，目的の情報がどの位置に含まれるかどうかという判断は章節のタイトルから判断されるのが一般的であると考えられる。

これらの既存の印刷物の利用調査の結果から，Dillonは全文データベースの設計に対し，(1) 目次，(2) 論文についての簡潔な情報（標題，著者名，抄録，選択可能な節見出し一覧，引用文献，論文の長さなど），(3) ブラウジング機能，(4) 自在に欲しい部分を印刷できる機能などが求められていると提言している。[20]

Ellis は，社会科学系の研究者を対象として，研究の過程における情報利用行動について研究している。この調査の結果，

①探索の開始

②引用の利用

③ブラウジング

④文献の差別化

⑤研究分野の監視

⑥抽出

という六つの特徴的な利用行動の過程があるという結論を出している。これらの過程の段階の中には，全文を読む段階だけでなく，文献の書誌事項や引用情報を利用する行動や，求める情報を含む文献の部分のみを選択し利用する行動が見られている。ここから，文献の部分のみを検索するという行動に対しての妥当性を引き出すことが可能である。[21]

Kercz は，学術文献の利用者を四つの立場に分類し，それぞれの読み方と検索要求を分析している。ここでの分類している利用者は，

①研究の管理者

②分野の専門家

③研究分野への新規参入者・隣接領域の研究者

④新しいことを学ぼうとしている人

となっている。

①のレビュー論文を書くレベルの研究の管理者は，書誌的な情報が必要であり，②の専門家は，背景知識を持っているために，論文の中心的な内容のみが必要で，論文全てに目を通すことはあまりないと考えられる。この①②の利用者は，自分の研究経験から，抄録やキーワードを対象として検索を行うシステムを有効に活用できるとしている。しかし③や④の利用者は背景知識を持たないために，書誌データベースにおいて検索式をうまく組み立てることは難しい。このような利用者には，「なぜ」「どのように」といった従来の自動検索システムでは表現できない方式や，文献のブラウジングが有効であると考えている。そこで，全文データベースにおける構造を利用して，検索することが有効であると述べている。また，この論文の構造を標準化する「メタレベル構造」を定義し，これを利用することにより，一層利用者が有効な検索ができると提案している。[22]

文献の構造に基づいて質問が設定されるという研究は，Allen も行っている。Allen は，“文献の利用者は，文献が van Dijk らの提示した上部構造（第1図）を持つことを期待している。そしてその上部構造を手がかりとして文献の内容を理解し想起している”と述べている。さらにこの文献の構造は，文献を探索する上でも利用されると考えている。[23]

また，神門は文献に記録された内容の特性を，主題領

Introduction
Setting
Time
Place
Literature
Purpose
Method
Results
Discussion

第1図 van Dijk の示した文献の上部構造

域や形態的な差異にとらわれず，共通の枠組みで捉えることを目的とし，日本語の医学，国文学など4つの領域の原著論文を対象として，その内部構造を分析している。そして，この分析の過程で，階層構造を持った構成要素カテゴリの体系を作成している。この構成要素カテゴリを用いた構造分析は文献の記録内容をとらえる共通の基盤を提供し，情報メディア研究やその生産・蓄積・利用などの諸側面，特に，全文データベースでの文献の一部を単位とした利用や検索の高度化に利用できるものと考えている。しかし，全文データベース検索にこれを用いる場合には，これらのカテゴリを自動付与することは不可欠なことであり，言語学的分析や，高度な自然言語処理の技術が必要となるものと考えられる。[24]

このように，学術雑誌の特性やその利用状態，また利用者特性や検索の特徴などにおける調査が幾つか行われているが，これらを全文データベースシステムの設計の際に考慮していくことによって，より利用しやすく，また検索効率の高い全文データベースシステムを構築していくことが可能である。

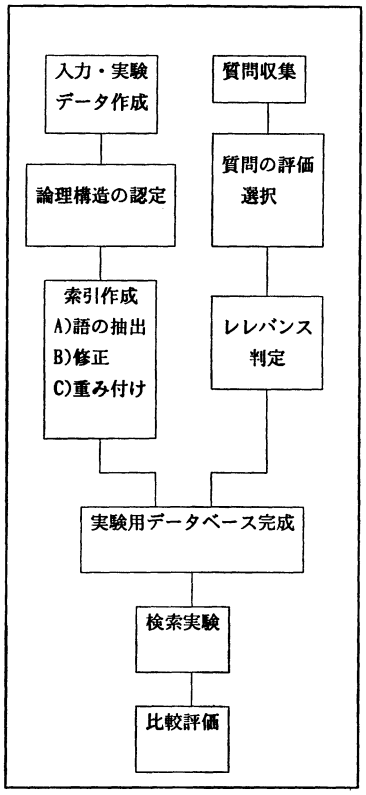
IV. 検索実験

A. 目的

従来の書誌データベースでは，一つの論文や記事を検索の単位としていた。そして，全文データベース検索においても，論文，記事全体が検索結果となっている。しかし，最初に挙げた全文データベースの特色と前節で示したような「論理構造」に着目した場合，検索する単位は，必ずしも全文に限る必要がない。全文データベースでは個々の章，節，あるいは段落，文を単位とした検索が可能である。

以上のような可能性を踏まえ，ここでは，「文」に対し一つ上の階層を構成しており，どのような全文データベースにも出現すると考えられる段落を単位とした検索を試みた。段落に着目するのは，論文などの利用方法，読み方を検討すれば，全文データベースの検索においては必ずしも文献全体ではなく，必要な情報が記述されている部分だけが提供されれば良いという場合がありうるためであり，また，段落が主題を表す最小の単位であると考えるためである。

検索方法としては，論文の論理構造を利用するとともに，各種の検索手法の適用の可能性を探った。特に雑誌論文では，一つの論文の中に多数の主題が記述されており，これらの主題を個別に検索対象として扱えるようにすることは有意義であると考えられる。なお，以下で対象とするのは，論文の全文であって，図表などは除いたテキストのみからなるデータベースである。



第2図　実験の手順

実験手順のフローチャートを第2図に示す。

B. 検索実験用データベース

検索対象となる文献集合は日本語の学術論文から選定した。論文の主題は情報検索，自然言語処理の分野に限定し，1972年から1992年までに出版された『情報処理学会論文誌』，『電子通信学会誌』，『情報処理』，『Library and Information Science』の4誌に掲載された原著論文を読み，選択した。この際の基準となっているのは，図表を含めて6ページ以上の文章があるものとし，レビュー論文は除いた。選択した文献数は全部で 49 件である。

選択した文献を OCR で読みとり，最終的な修正を人手により行って，実験用データベースを作成した。作成した実験データベースは，図，表を除く全文（但し，図

表タイトルを含む）である。なお，文章中には数式も含まれている。最終的にこのデータに対し，論理構造を認識するための SGML タグを付与した。タグの付与は一部の単純な部分について自動で行ったが，そのほかのタグは OCR 読みとりの結果の修正と同時に人手により行っている。

SGML の規格については，その表記方法や，タグについても様々な案が出されている。その中で，本実験においては，「ISO 10283-1993 電子出版における文献の様式，構成要素，告知」を適用した。これは，図書，論文，逐次刊行物等の資料タイプに適用されるものである。これを用いることにより，資料を構成するタイトル，著者，本文などの資料要素が識別可能となる。このほかに，全資料タイプ共通で用いられる，段落や書誌事項，引用，図表などの資料要素については，「共通資料要素」としてのリストが挙げられている。これらのタグについては，文章中における特定の箇所を参照しその内容を呼び出す機能を持っている。参照は，資料が処理される時点で実体と置き換えられる。参照の例としては，脚注，図，表などが挙げられる。

C. 質問収集

検索者が，どのような検索要求を持ち，また部分を対象とするような検索質問が実際に表われるかどうかを知るために，検索質問として実験者本人の質問だけではなく，一般の研究者，大学院生を対象として調査，収集を行った。

各被験者にデータベース中の任意の2～6文献から，特に論文の部分を構成する個々の主題を対象とするような質問を，10～50字程度の文章で記述するよう求めた。この時，検索質問は被験者自身の語で表現し，さらに検索質問に取り入れることを望む語を含めるよう依頼した。

さらに，被験者が記述した質問式がどの部分と適合しているのかという記述を，頁，章，節，段落などで提示することを求めた。この提示結果を参考として，実験者が検索語の選定と重み付けを行う検索式の作成，質問に対する 49 件の論文を対象としてレレバンス判定を行った。なお，収集した質問は43問であったが，レレバンス判定を行ったのは，このうちの8問についてである。[25)]

被験者は，慶応義塾大学の自然言語処理研究に携わる理工学部大学院生，図書館・情報学科の教員，大学院生および学部生から計9名からなり，収集した質問数は43問である。

D. レレバンス判定

一つの論文中においても様々な主題についての記述があり，ある部分についてはレレバントであるとしてもそのほかの部分は不要と言った場合もある。この実験では，部分を対象としたレレバンス判定が必要とされる。そこで，検索を行う前に，それぞれの論文の段落毎にレレバント，ノンレレバントの二値でレレバンス判定を行った。

論文の段落を単位としてレレバンス判定を行うことには，その単位が妥当であるかという問題がある。レレバントな部分は，(1) 章や節，(2) 二つ以上の段落，あるいは，(3) 段落中の一文といったことが考えられる。ただし，(1) 章や節と ②二つ以上の段落については段落を単位としてレレバンス判定を行うことが可能である。③段落中の一文についてはレレバンス判定をもっと細かい部分で行う方がよいという見解も存在する。しかし，実際にこの単位でレレバンス判定を行うことは非常に困難である。そこで本実験では，この一文単位のレレバントも存在すると考えた上で，一文のみがレレバントである場合については，その文を含む段落をレレバントであると判断した。

この他に問題となった点としては，「A については，B（節）でのべる」といった指示的な文章における A を，レレバントの対象とするかどうかである。このような文章は随所に見られるが，本実験においては，部分検索が一次情報を得られるところに利用価値があると考えるため，これについてはノンレレバントとして評価した。しかし，「このシステムでは A が必要とされる」という段落の後で，「A は…。」と説明する場合の A については，前者の段落も A についての説明がなされていると考え，これはレレバントな段落の対象とした。

国語辞典や，百科辞典において見出しの項目で検索を行う場合，その項目について説明している部分のみを提供すればよく，前後の項目との境界は明確である。同様に従来の抄録データベースによる検索実験では，それぞれ独立している各文献単位でレレバンス評価を行えば良い。しかし，全文データベースの場合は事情が異なっている，文献の構成要素は，それぞれが独立し，完結している情報ではなく，前後，相互のつながりがあるため，部分に分けてしまうとレレバント，ノンレレバントの境界が明瞭ではなくなる場合があるからである。この前後関係のどのレベルの情報までをレレバントと考えるかによって，検索結果の評価の際に大きな差が現れる。

このように，段落を対象として一段落毎にレレバンス判定を行うにあたっては何らかの基準，指針が必要である。本研究では，「前後の段落との関係は考慮せず，その段落単体で検索結果に答える内容を明示しているかどうか」，という点を念頭に置いてレレバンス判定を行った。

E. キーワードの抽出

キーワードの抽出処理のフローチャートを，第3図に示す。この順に沿って，以下で説明する。なお，キーワードの重み付けについては次節で説明する。

まず，カタカナ語，アルファベット，漢字の熟語（2字以上の漢字の場合），これを切り出す。また，漢字の熟語については，4字以上の漢字については，2字ずつ最初から切り離した状態で抽出した。

次に，接頭辞，接尾辞，助数詞を，最初のものから取り除いた。この辞書は，実際の索引を修正する上で本実験環境用に作成した。なお，この辞書において切り離されるうえで不備があると考えられる点もあるため，完全自動化ではなく人手により確認し，修正を行っている。

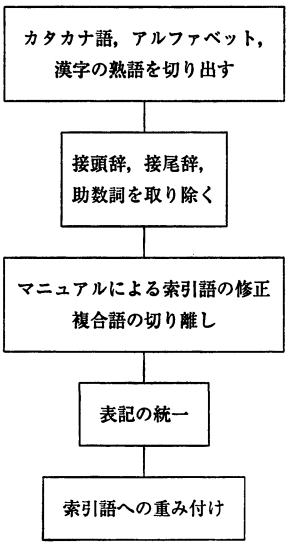
また，キーワードの修正段階において，以下の2点の表記の揺れを統一修正をした。

（左から右へ修正）

①英語のカナ文字表記にしたもの

例）　カテゴリ→カテゴリー

　　パターン→パタン



第3図　キーワードの抽出処理のフローチャート

　　デジタル→ディジタル

　　シンタックス→シンタクス

　　セマンティクス→セマンティックス

②難漢字の表記のかな表記

例）　語い→語彙　　ら列→羅列　　網ら的→網羅的

また，今後は指示語の照応語への変換，省略語の補足なども考えていく必要がある。特に，同一文，同一段落

	α	β	γ	σ	ε
[タイトル]					
情報					
検索					
[章]					
日本語					
検索					
実験					
[パラグラフ]			日本語 0.1	日本語 0.4	日本語 0.2
データ	データ 0.1	データ 0.1	データ 0.1	情報 0.15	情報 0.15
実験	実験 0.1	実験 0.1	実験 0.1	検索 0.8	検索 0.8
情報	情報 0.1	情報 0.1	情報 0.1	実験 0.5	実験 0.5
抽出	抽出 0.1	抽出 0.1	抽出 0.1	データ 0.1	データ 0.1
検索	検索 0.1	検索 0.2	検索 0.1	抽出 0.1	抽出 0.1
検索			結果 0.1	結果 0.3	結果 0.3
[表タイトル]					
検索					
結果					

第4図　重み付けしたデータの例

第1表 索引語の抽出と重み付けの方針

論文中の部分	索引抽出方針と留意点	段落検索による重み付けの方法 ($\delta \cdot \varepsilon$)
論文タイトル	当該論文中において，中心主題となるキーワードが出現していると考えられるため，本文中に出現した場合に，そのキーワードに重み付けを行う等，重要語句であると認定する。	本文中，章，節タイトル等すべてに出現している語句に対し，その語句の重み付けを1.5倍する。
抄録	本文の一部とも考えられるが，今回は検索，索引語抽出の対象とはしない。	
章タイトル 節タイトル	章・節タイトル中に現れるキーワードは，論文タイモルに現れるものよりも更に直接的にその下部構造と関わっていると考え γ，δ，ε ではその語を各段落のキーワードとして付与する。	各段落毎に重みを δ では 0.4，ε では0.2として付与。 ＊ 章，およびその下部構造の節，両者に同じ語が現れる際には二度目以降の重みを半分として付与する。
強調語	[ ] に入っている語，アンダーラインのある語についても重み付けを行う際に考慮する。また，下部構造がある場合には章タイトル等と同様の処理を行う。	重み付け 0.4，下部構造に対する処理は章タイトルと同じ
箇条書き語	前後の関連する（関連の強い方の段落）に取り入れ，独立段落とはしない。	出現頻度一回につき，0.2で重み付けを行う。
例	抽出対象とはしない。	
段落文中	原則的に行替えと，一文字下げられているものを段落とする。(但し，箇条書き，アルゴリズム，公式等を例外とする)	出現頻度一回につき，0.1で重み付けを行う。
表タイトル 図タイトル	参照のある段落に（数回出現するものもある），キーワードとして取り込むようにし，図表は重要な情報源であると考えられるため，ここから抽出さるキーワードへの重み付けは高くする。	参照のある段落に図表タイトル中のキーワードを0.3で与える。 ＊ 各段落内で何回図表が出現しても，一回と数える。
図表脚注	抽出対象とはしない。	
図表内容	抽出対象とはしない。	
公式	公式中に含まれるキーワードは，一般の段落中に現れるキーワードと同様に，取り入れている。	出現頻度一回につき 0.1
アルゴリズム	公式と同様，段落中に現れるキーワードと同じ処理を行う。	出現頻度一回につき 0,1
本文脚注	参照のある部分に組み込む。	出現頻度一回につき 0.1
引用文献タイトル	抽出対象とはしない。	
付録タイトル	参照のある部分へ組み込み，図表タイトルと同様の処理を行う。	参照のある段落にその付録タイトル中のキーワードを0.3で与える。
付録	付録内容からは，キーワードとしてはとらない。(例文的なもの，付表となっているものであったため)	
謝辞	段落の扱いとおなじ（1段落と考える）	出現頻度一回につき 0.1

中においては，指示語，省略語の出現頻度が高いと考えられ，出現頻度による重み付けの上で重要な概念の抽出が効果的に行われるかという問題と深く関わっている。ただし，本研究ではこの補足検討は行っていない。

F. キーワードへの重み付けと検索への利用

上記の処理の結果，段落毎に抽出されたキーワードに対し，重み付けの可能性について検討を行った。部分を対象として検索を行う際，より適合度の高い段落から検

索されることが望ましい。そこで，段落中における重要度の高い語については重みを高くすることで検索システムの高度化を図る実験を行った。

重み付けの根拠としては，手作業で索引作成を行う場合の留意点を参考とした。例えば，「ISO 5963 索引作成」では，キーワード抽出の際に参照する箇所として，標題，抄録，目次，序文，章や段落の最初の部分，結論，図表と図表名，太字・イタリックや下線が引かれた語などを挙げている。また，医学分野のデータベースであるMEDLARS では，(1) 標題，(2) 文献の目的が示されている箇所，(3) 本文中の章節の見出し語，太字やイタリックで表されている語，図表，(4) 概要記述部分，(5) 抄録，(6) 注・引用文献を挙げている。[3)]

これらを検討し，タイトル中のキーワードに加え，章，節のタイトル，図表タイトルなどについては，本文中の文章に出現した語よりも高い重みを付与する方針を立てた。また，繰り返して出現している語については語の重みを高くした。ここで用いている重み付け方法は，出現頻度と出現位置による統計的手法を用いたものである。

比較評価のために，以下の α から ε の重み付けによる索引を作成した。

α: 段落中の語を切り出しただけの索引

β: 段落中の語 (α) に出現頻度による情報を加えた索引

γ: 段落中の語 (α) に章，節タイトル，図表タイトルを加えた索引

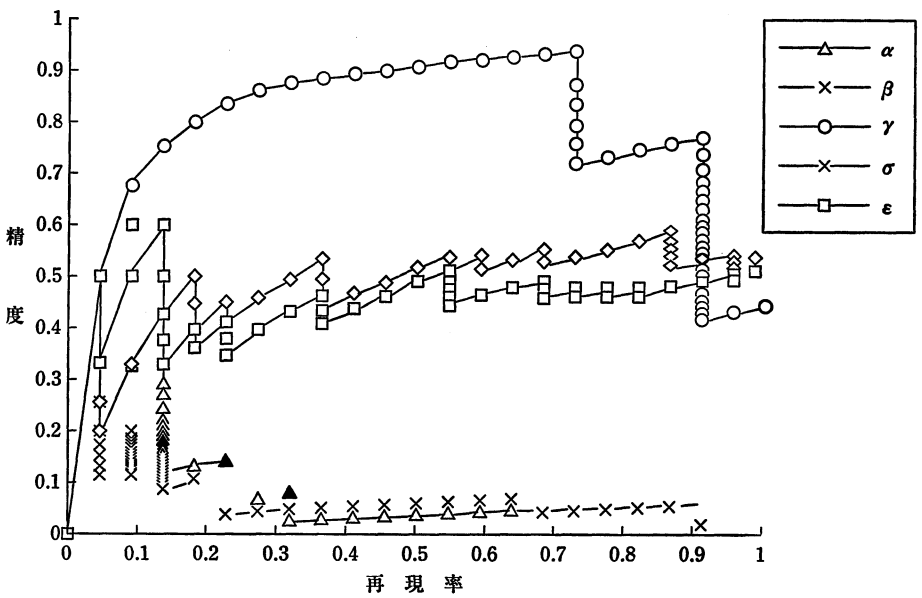
δ: 段落中の語，章・節・図表タイトル (γ) に出現頻度，出現位置の情報を考慮した重み付けを行った索引

ε: δ の章，節タイトルの重みを修正した索引

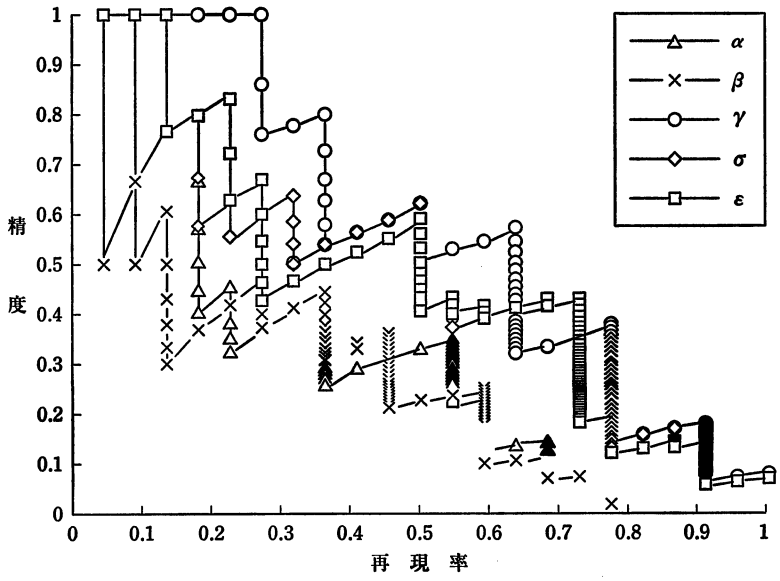
なお，試験的な実験の結果から，章，節のタイトルが及ぼす影響が大き過ぎると考えられたため，後に δ に修正を加えた。これが ε である。重みの加算法の例を第 4 図に示した。この図では，タイトル，章タイトル，段落，表タイトル中から自動的に切り出された語をもとに，前記の α〜ε のそれぞれで，どのような語がそれぞれの段落に付与され，重み付けが行なわれたかを表している。

全文データベースに対し SGML のタグを付与して認定した論理構造の要素の種類と，その要素別の重み付け方法を第 1 表に示す。これにより付与される重みを段落毎に加算し，最終的な索引とした。

また，検索結果については，質問と各段落のキーワードの類似度を計算して適合度順出力を行った。この際に用いたアルゴリズムは，SMART システムで用いられているベクトル型モデルである。[26)]



第 5 図　検索結果 (1)



第6図 検索結果(2)

G. 結果

検索の評価は，適合度順に出力を行った結果に対し1段落毎に再現率と精度を計算している。このグラフを，第5図と第6図に示す。

第5図，第6図ともに文中の語の出現頻度を考慮した β と α ではほとんど差は現れない。ほかの質問においても同様な傾向が見られ，絶対頻度を用いて重み付けを行った効果は表れなかった。これについては，より大規模なデータベースで実験を行い，レレバントなパラグラフがもっと多数ある場合には，何らかの差は見られると考えられる。しかし，本実験における小規模なデータベースでは，α と β の出力段落順位はほとんど同じであった。

しかし，章・節・図表タイトルを含めている γ, δ, ε における検索は，他のどの質問においても α, β を上回った結果となった。これは，質問式に含めた語が章・節・図表タイトル中に含まれたためであるという理由はあるが，これらの語が段落の文章中には表れていないという結果も明らかになった。

V. 考　　察

以上の全文データベースの検索実験では，章，節タイトルを段落に含めて検索を行う方式が，最もよい検索結果を示している。一方，重み付けを行った β をはじめ，δ, ε の結果は良好な結果を示さなかった。これは，別のパラメータや，重み付けを行い検討を重ねる必要があると考えられる。特にここでは，重み付けの方法として，単純に加算する方式を採用したが，段落の長さが算出される適合度に与える影響が問題点として挙げられる。これは，長い段落であれば，算出される適合度が増加し，上位に検索される可能性が高くなるということである。この点については，(1) 段落の総語数との比をとる，(2) 段落の異なり語数との比をとる，(3) これら (1) や (2) に該当段落を含んでいる文献について，全語数を考慮して比をとるなどの正規化を行う必要がある。

部分テキストを検索するにあたって，論文中の章，節などのタイトルを各段落の検索語として取り込み，検索を行うことは有効であることが分かった。この理由として，論文の章，節のタイトル中の語は，省略，または指示語の形で記述される場合が多いため，段落への強制付与の方法が有効であると考えられる。

また，タイトル中の語を盛り込むことにより，第1図で示した，van Dijk らの述べている論文の上部構造に基づく質問を受け入れることが容易となることが考えられる。

これに基づくと，章や節のタイトルにこれらの“目

的", "方法", "結果", "結論" といった語が含まれている場合，この上部構造を想定する検索質問の設定が可能となる。

最後に，章，節タイトルなど，重要であると考えられる部分に重み付けを行って検索することにより，検索効率を高めることができなかった理由を考えていく必要がある。この理由としてはまず，レバンス判定が適切であるかという前提としての問題と，重み付けの方法である各構成要素についての重み付けパラメータが適切ではない場合が考えられる。また，検索システム側の問題だけではなく，論理構造を利用した検索の場合，論文を記述する著者が適切な論理構成で書いているかという問題も存在している。

VI. おわりに

抄録や索引を作成する際には，その文献の主要な概念を選択するという過程がある。そのために，文献中の重要な概念については抄録中に現れるが，この概念を補足する概念や関連する概念などで，索引・抄録作成者が注意を払わなかったものについては検索の対象とならなくなる。この過程ですでにモレが生じている。一方このモレは，全文中の語を対象とする場合には生じるものではない。全文を対象として検索することの利点としては，この索引の網羅性が高いことである。

抄録などに比べ全文からはより多くの主題概念を表す語が提供されるために，当然より多くの文献が出力される。重要な主題概念である語は抄録中に出現しているとしても，1 段落，1 文などのごく一部分で表現されるような主題は，書誌データベースで抽出されない可能性が高い。しかし，これらの一部分に表現された主題が不要なものであるというわけではない。

このような理由から，全文データベース検索では，レレバントであると判断される部分が検索されれば十分である場合が想定される。または，この部分を判断対象として全文，もしくは前後を含む拡大部分を入手し，その文献の適合性を検討する方法が有効であると考えられる。

この部分の単位が，利用者にとって必要な情報を得られる最適な量で，個々の文献を蓄積させることは，個々の文献の主題や叙述方法の違いにより不可能なことである。このために，検索される部分の単位が SGML を用いて認識されるような，文，段落，章，節などの論理構造を手がかりとすることは実用的なものであると考えられる。

本研究では全文データベースにおいて，部分を対象として検索する方法の有効性を検討し，段落を単位として検索，提供を行うことにより，レレバントな情報が検索されることを明らかにした。また，その際に，論理構造を認定することで取り入れることができる章，節などのタイトル中の語を用いることで，より検索効率が向上することを実験により確かめた。

しかしながら，段落だけで提供されるのでは，前後のつながりや背景情報などがないために，検索結果の有効性を判断することが困難である。そこで，検索された各段落に対し，前後の段落や，その章，節などの論理構造をたどって，自由にブラウジングできるようにすることが必要であると考えられる。

すなわち，全文データベースの部分を対象とした検索においては，文献中の部分を各主題毎にまとめ，ノイズとなる部分を取り除き，適合する部分のみを提供することができるようにしなければならないが，論文の場合には，前後の「部分」とのつながりがなければその部分を理解することができないという問題がある。そのため，検索された部分を手がかりとして，その関連する部分や上位構造をたどり，出力できるような，柔軟性の高いインタフェースを構築して補完する必要がある。

謝辞　本研究を進めていく上で，慶應義塾大学図書館情報学科上田修一教授からは終始きめ細かくご指導頂きました。またこの実験を行うにあたって多くの方々から，実験データ，検索質問を提供して頂きました。ここに記して，心より御礼申し上げます。[27]

1) Blair, D. C.; Maron M. E. "An evaluation of retrieval effectiveness for a full-text document-retrieval system". Communications of the ACM. Vol. 28, No. 3, p. 289-299 (1985)
2) Blair D. C.; Maron. M. E. "Full-text Information retrieval further analysis and clarification". Information Processing & Management. Vol. 26, No. 3, p. 437-447 (1990).
3) 細野公男編. "2.2 主題索引作業". 情報検索. 東京, 雄山閣, 1991, p. 44-51. (講座: 図書館の理論と実際　第5巻.)
4) Salton, G. Introduction to modern information retrieval. New York, McGraw-Hill, 1983, 448p.
5) Salton, G.; Buckley, C. "Term-weighting approaches in automatic text retrieval". Information Processing & Management. Vol. 24, No.

5, p. 513-523 (1988)
6) 長尾真. “日本語文献における重要語の自動抽出”. 情報処理. Vol. 17, No. 2, p. 110-117 (1976)
7) 横井俊夫. “特集: 新しいデータ・新しい研究, 電子化辞書とテキストデータベース”. 日本語学, Vol. 10, No. 8, p. 78-85 (1991)
8) 吉村賢治. “日本語科学技術文における専門用語の自動抽出システム”. 情報処理学会論文誌. Vol. 27, No. 1, p. 33-40 (1986)
9) 小西修. “自動構築型知識に基づく専門用語形成システム”. 情報処理学会論文誌. Vol. 30, No. 2, p. 179-189 (1989)
10) 絹川博之; 田中和明; 池上信男. “日本語情報検索システムにおけるキーワード自動抽出”. 日立評論. Vol. 64, No. 5, p. 75-79 (1982)
11) 木本晴夫. “日本語新聞記事からのキーワード自動抽出と重要度評価”. 電子情報通信学会論文誌. D-I, Vol. J74-D-I, No. 8, p. 556-566 (1991)
12) 中村隆宏; 秋田昌幸. “日本語の重要度評価について”. 1990 年度人工知能学会全国大会 (第 4 回) 論文集. p. 293-296 (1990)
13) Hickey, T. B. “Using SGML and Tex for an interactive chemical encyclopedia”. Proceedings of the 10 th national online meeting. Medford, NJ, 1989-5. Learned Information, New York, 1989, p. 187-195.
14) 影浦峡; 根岸正光他. “文献の論理構造を考慮した全文検索システム”. 学術情報センター紀要. No. 3, p. 49-58 (1989)
15) 石塚英弘ほか. “日本化学会欧文誌の SGML 形式全文データベースの構築・印刷そして検索”. 情報学基礎研究報告. No. 29, p. 1-8 (1993)
16) 長尾高明. “特集: 文章と談話, 文章と段落”. 日本語学. Vol. 11, No. 4, p. 26-32 (1992)
17) 森岡健二. 文章構成法: 文章の診断と治療. 東京, 至文堂, 1981, 546 p.
18) 佐久間まゆみ. “特集: 文章と談話, 文章と文: 段の文脈の統括”. 日本語学. Vol. 11, No. 4, p. 41-48 (1992)
19) 永野賢. 文章論総説: 文法論的考察. 東京, 朝倉書店, 1986, 379 p.
20) Dillon, A.; Richardson, J.; McKnight, C. “Towords the development of a full-text, searchable database: Implications from a study of journal usage”. British Journal of Academic Librarianship. Vol. 3, No. 1, p. 37-48 (1988)
21) Ellis, D. “A behavioural model for information retrieval system design”. Journal of Information Science. Vol. 15, No. 4, p. 237-247 (1989)
22) Kercz, Joost G. “Retorical structure of scientific articles: the case for argumentational analysis in information retrieval”. Journal of Documentation. Vol. 47, No. 4, p. 354-372 (1991)
23) Allen, B. “Text structures and the user-intermediary interaction”. RQ. Vol. 27, No. 4, p. 535-541 (1988)
24) 神門典子. 構成要素カテゴリーを用いた情報メディアの内部構造分析の試み: 原著論文を例として. 東京, 慶應義塾大学, 1991, 237 p. 修士論文.
25) 質問を収集した時点で部分を対象とした検索質問を記述することに対しての疑問が生じた. まず, ‘～の実験例’といった場合には, 方法, 実験, 結果, 考察まですべて含められることになり, 最終的に全文をレレバントであるとなってしまう. またこのような検索質問がほとんどであり, そのような場合には部分を対象としてレレバンス判定はできないうえ, 検索することはかえって計算量や時間を増加させるのではないか, といった意見が出された. 実際に, 今回収集した検索質問からも, かなり部分的にレレバントであると判断しにくい質問がいくつか見受けられた. このような場合についても, 考慮しなければならないが, 今回採用した検索質問は比較的, 特定的な部分に限ってレレバントであると認識しやすいものとした.
26) SMART 検索システムにおけるマッチングアルゴリズムは以下のようになる.

$$\mathrm{COS}(D_i, Q_j) = \frac{\sum_{k=1}^{t} (\mathrm{TERM}_{ik} \cdot \mathrm{QTERM}_{jk})}{\sqrt{\sum_{k=1}^{t} (\mathrm{TERM}_{ik})^2 \cdot \sum_{k=1}^{t} (\mathrm{QTERM}_{jk})^2}}$$

D_i=文献 i
TERM=文献 i に含まれる語の重み
Q_j=質問 j
QTERM=質問 j に含まれる語の重み
t=文献群全体に含まれるすべての語の数
27) 本研究は, 野末道子. 平成 4 年度慶應義塾大学大学院文学研究科図書館・情報学専攻 修士論文“論文の論理構造を利用した全文データベースの検索”に基づいている.