

論文段落を対象とした 日本語全文データベースの検索

鉄道総合技術研究所 野末道子

慶應義塾大学文学部 上田修一

全文データベースでは再現率は向上するが、精度は著しく低下するなど、検索の面でいくつかの困難があることが指摘されている。全文データベースの検索では、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文献の一部分だけが提供されればよいという場合が考えられる。そこで、部分テキストの検索の有効性について実験を行った。具体的には、文章の段落を中心として、段落を一つの単位として、さらにいわゆる論理構造に基づいた情報を用いた5種類の索引法を施し、検索実験を試みた。実験対象は、日本語の情報検索や自然言語処理関連の論文49件で、あらかじめ収集した検索質問によってレlevance判定を行っている。この実験の結果、段落を用いた検索の有効性と、その際に章、節タイトル中の語を含めて検索を行えば検索効率が高まることが確かめられた。

Full-text database retrieval using paragraphs

Michiko NOZUE

Railway Technical Research Institute.

2-8-38, Hikari-cho, Kokubunji-shi, Tokyo 183, Japan

Shuichi UEDA

Library and Information Science, Keio University.

2-15-45, Mita, Minato-ku, Tokyo 108, Japan

Because recall is higher than bibliographic databases, and precision is so lower, full-text database is difficult to retrieve. We examined the retrieval effectiveness of full-text database retrieval by using paragraphs of individual documents. Sample documents are 49 articles in Japanese. The higher precision and recall was shown by using the words in chapter titles or section headings to retrieve the relevant paragraphs.

1. はじめに

従来、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の研究がなされてきた。全文データベースの発展にともなって、日本語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大きな点で異なっている。一つは全文が検索対象となるという点である。書誌データベースでは、主題探索の際の検索対象は、標題、抄録、それに付与された索引語群、あるいは引用文献などであった。もう一つの相違は、検索と同時に全文を入手できるという点である。書誌データベースを検索した後に、別個の手順によって適合文献を入手する必要があった。

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みなされることが多い。

- ①文章中の語および、図、表、写真等も検索が可能な全文データベース
- ②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、写真等の画像へのアクセスポイントの付与、および出力の問題は未解決となっている。③の例としては、例えば「Business Periodicals On disc」や「ADONIS」のようにCD-ROMで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パッケージがある。

2. 全文データベース検索の課題

2.1 全文データベースの論理構造の利用

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あげることができるが、これらは、文、段落、あるいは章、節という単位で分割ができ、こうした構造は一般的に「論理構造」と呼ばれている。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。

2.2 全文データベースの検索手法の開発

さて、商用データベース・サービスの多くは、全文データベースの検索手法として書誌データベースの検索手法を踏襲している。例えば、わが国の新聞記事データベースでは、全文を対象として付与された索引語と全文から自動抽出した語のインバーテッドファイルを作り、ブール演算子等を使用して検索している。一方、サーバ・クライアント方式で全文データベースの高速検索を意図した検索システムも開発されており、例えば新日鉄の「NSEARCH」は、文字列のパターン化を行なって、検索速度の高速化と検索の柔軟性を高めている。このように全文データベースの検索は、ファイルの価格の低下や、プロセッサの処理速度の高速化によって支えられていると言える。

しかし、これらのいずれも全文中の文字列をそのまま利用するものであり、全文データベースのテキストの構造やこれまで考案されてきた検索手法はさほど考慮されてはいない。

2.3 段落を単位とした検索の可能性

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、最初にあげた全文データベースの特色と2.1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。特に雑誌論文では、一つの論文の中に多数の事実が記述されており、これらを個別に検索対象として扱うことが可能である。

以上のような課題を踏まえ、ここでは、「文」の一つ上の階層となっており、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、また論文の論理構造を利用し、さらに各種の検索手法の適用の可能性を探った。

なお、以下で対象とするのは、論文の全文であり、図表等は除いたテキストのみからなるデータベースである。

2. 情報検索手法の発展

検索手法については、商用オンライン検索の主流である論理演算によるものだけではなく、多様な方法が提案されており、実験環境ではかなりの成果が得られている。最初にこれらの索引方法、および検索モデルについて概観する。

2.1 索引法

抽出索引法の中では、確率・統計的手法を用いて索引語を抽出する方法、文章を構文解析し適切な索引語を決定する方法などが主流となっている。コーネル大学で開発されたSaltonのSMARTシステムが代表的である。^[1]

一方日本語文献の索引抽出処理には、英語文献を処理する場合とは異なった、日本語特有の問題がある。このうち、

- ①文が単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漢字は表意文字であり、熟語、複合語を作る造語性が高い

④外来語は片仮名で表わされ、表記のゆれを生じ易い。

などが大きな問題となっている。

全文を対象として索引語抽出実験を行っている例としては絹川^[2]、木本^[3]、長尾^[4]らの実験がある。しかし、上記のような理由から、全文を対象とした日本語の索引語抽出には、数多くの研究課題が残されている。

2.2 検索モデル

商用データベース・サービスで用いられている論理型の検索モデルの他に、必ずしも指定したキーワードをすべて含む文献を検索するのではなく、部分的に合致している文献を検索する、ベクトル、ファジィ、確率型といった検索モデルが存在している。^[5]これらの部分一致型のモデルでは、キーワードをどれだけ含んでいるかという状況による適合度順の出力、および自動的な検索質問のフィードバックが可能となっている。これらのモデルの特徴は表1のように要約される。

表1 主要な検索モデル

	論理型	ベクトル型	確率型 (決定理論型)	ファジィ型 (拡張論理型)
特徴	AND、OR、NOTといったブール演算子を用いて検索語と索引語を照合する。	検索質問と文献を語のベクトルとして表現し、類似度を計算する。	特定の語を含む文献が適合、不適合となる確率を考慮した関数を用いて検索する。基本的にフィードバックを前提とする。	検索質問と文献中の語をメンバーシップ関数で表現し、ブール演算を行う。

4. 学術論文の特徴と論理構造

以下では学術論文の全文データベースを対象とする。

4.1 論文の分割と段落の特徴

文章では、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、我々がものを理解するときには分割と総合という思考上の働きがあることの反映と考えられている。文章では著者によって「段落」という目に見える形で分割が行われている。森岡は、段落とは文章全体を適当な部分に分割する方法、主題を支える論点や材料を述べる手段、独立して文章全体の一局面である小主題を述べる手段であるとも

なしている。^[6]

しかし、実際の段落には、トピックセンテンスの欠如も多くみられ、意味的まとまりとして段落分割がなされていない場合も数多く存在する。特に、文章の書き方についての教育やその必要性が比較的軽視されている日本語の国語教育の現状にあっては、段落分割を正しく行うことがあまり徹底されているとは言い難い。また、客観的に第三者が考える段落分割と著者の認識する分割が異なる場合、形式段落と意味段落の区別など、段落がそのまま小主題を述べているものとは断定しにくい状況も考えられる。そのために、形式段落を意味的まとまりとして意義のないものと判断し、意味段落もしくは文段を認定する議論が文章論などではなされてきた。

しかし、こうした段落の認識は個々の読者の認識に基づくものであり、確定的なものではない。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」と考えることは妥当であろう。^[7]

4.2 論理構造

文章の長さによって、段落分割はさらに、章、節、項等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。この階層構造によれば、章、節、段落、文、単語の順に従って、主題がその細部へと展開していく様子が見られる。

しかしこれらの論理構造は、人間が目で見ても認識することは容易であるが、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り、その具体例としてSGMLタグ等が設定されている。SGMLにより記述されたデータベースは、文書の各論理要素を、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ている。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みられる。米国OCLCの化学百科辞典はその一例であり、この百科辞典には、図表、テキスト、数式などが含まれ、SGMLを利用して記述することによりページや参照を自由に検索できるようになっている。^[8] また学術情報センターにおいても、SGMLを用いた電子化が進められており、国内の学会誌製作にも利用されはじめている。^[9]

5. 検索実験

論文の特徴である論理構造を利用して、部分を対象とした検索の実験を行う。実験手順のフローチャートを図1に示す。

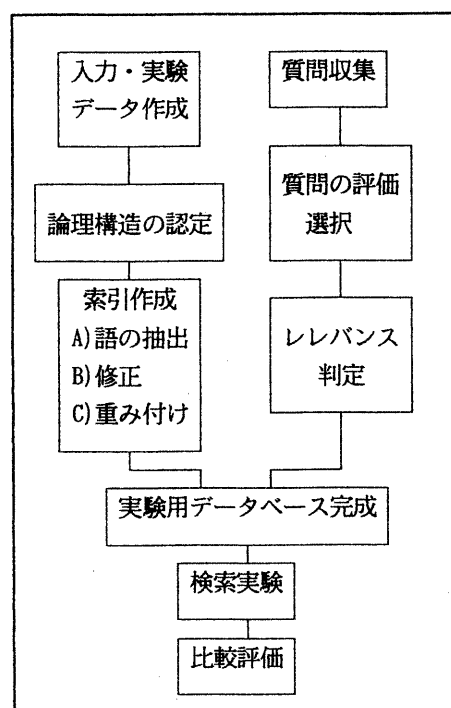


図1 実験の手順

5.1 検索実験用データベース

本実験を行うにあたり、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』、『電子通信学会論文誌』、『情報処理』、『Library and Information Science』の4誌に掲載された原論文を読み、選択した。データベースを構成する文献数は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、語を抽出することが可能な数式等も含まれている。またこのデータに対し、SGMLタグ（ISO DIS 12083を使用）を付与した。

5.2 質問収集

質問者が、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、大学院生などを対象として調査、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、検索質問は質問者自身の言葉で表現し、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、質問者が記述した質問がどの部分と適合しているのかという記述を、頁、章、節、段落等で提示することを求めた。検索語の選定と重み付けを行う検索式の作成、レlevance判定は実験者が行ったが、この提示結果を参考とした。

質問者は、慶大の理工学部、文学部の教員、大学院生等9名、収集した質問数は43問である。

5.3 検索比較実験

段落の検索を行うにあたって、段落中出现する語の利用方法の有効性を5種類の索引を作成し、比較した。この5種類の索引方法の根拠として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、索引語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句などをあげている。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・引用文献をあげている。^[10]

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、図表タイトル等については、本文中の文章に出現した語句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句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IからVまでの重みづけによる索引を作成した。

I : パラグラフ中の語を切りだしただけの索引

II : Iに出現頻度による情報を加えた索引

III : Iに章、節タイトル、図表タイトルを加えた索引

IV : III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

V : IVの章、節タイトルの重みパラメータを修正した索引

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いと考えられたため、後からIVに修正を加えVを加えた。重みの加算法の例を図2に示した。この図では、タイトル、章、パラグラフ、表タイトル中から自動的に切り出された語をもとに、上記のI～Vのそれぞれで、どのような語がそれぞれのパラグラフに付与され、重み付けされるかを表している。

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表2に示す。これにより付与される重みをパラグラフ毎に加算し、最終的な索引とした。

また検索は、各段落の類似度を計算して適合度順出力を行うベクトル型モデルを用いて行った。

	I	II	III	IV	V
[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					

図2 重みづけしたデータの例

5.4 結果

検索の評価は、適合度順に出力を行った結果に対しパラグラフ毎に再現率と精度を計算している。このグラフの例を、図3、4に示す。この結果、章、節タイトルを段落に含めて検索を行ったⅢが、よい検索結果となっている。

しかし重み付けを行っているⅡをはじめ、Ⅳ、Ⅴ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。とくにこの中で、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、パラグラフの長さが算出される適合度を与える影響が挙げられる。これは、長いパラグラフであれば、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、段落の長さによる正規化を行う必要がある。

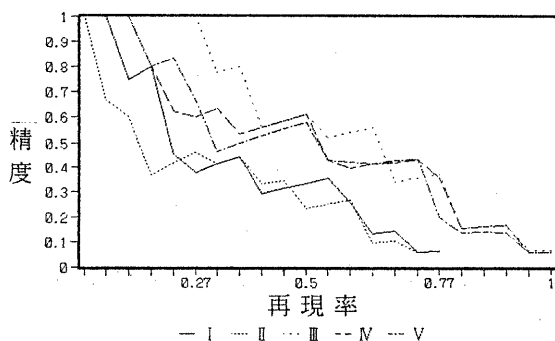


図3 検索効率 (1)

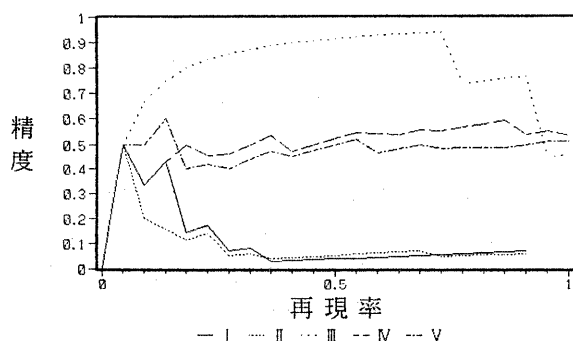


図4 検索効率 (2)

表2 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 (Ⅳ・Ⅴ)
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、論文タイト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Ⅲ、Ⅳ、Ⅴ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Ⅳでは0.4、Ⅴ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、ここから抽出す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

6. 全文データベースの検索手法と提供方法

部分を単位とする場合には、段落を単位として検索、提供を行うことにより、レバントな情報が検索されることが分かった。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句を用いることでより検索効率が向上することが実験により確かめられた。

しかし、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、検索結果が有効であるかという判断が困難であると思われる。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、自由にブラウズできるようにすることが必要であると考えられる。

今後は、実際の利用者がどのような検索質問を設定し検索を行うかを調査する必要がある。そして、それらの利用者が、実際に段落が提供されることによって必要な情報が得られるかどうか、また、どのような提供方法が求められるか検討する予定である。

謝辞 この実験を行うにあたって、実験データ、検索質問を提供して下さった皆様に御礼申し上げます。

- [1] Salton, G. Introduction to modern information retrieval. New York, McGraw-Hill, 1983, 448p.

- [2] 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)

- [3] 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)
- [4] 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)
- [5] Belkin, N.J.; Croft, W.B. Retrieval techniques. Annual Review of Information Science and Technology. Vol.22, p.109-130 (1987)
- [6] 森岡健二. 文章構成法: 文章の診断と治療. 東京, 至文堂, 1981, 546p.
- [7] 永野賢. 文章論総説: 文法論的考察. 東京, 朝倉書店, 1986, 379p.
- [8] Hickey, T.B. Using SGML and tex for an interactive chemical encyclopedia. Proceedings of the 10th national online meeting. Medford, NJ, 1989-5. Learned Information, New York, 1989, p.187-195.
- [9] 影浦峽; 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, p49-58 (1989)
- [10] 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.)