情 報 学 基 礎 31－6
（１９９３．９．１４）

文書の構造解析に基づく文書情報検索

三池誠司　小野顕司　住田一男

(株)東芝 研究開発センター

全文検索システムのための文書構造解析による検索方法とそのインプリメンテーションについて述べる。本検索方法では、文書構造の解析により、検索語が含まれている文の情報の種類を識別して検索し提示する。文書構造の解析は、文書書式、文脈構造および文役割の解析からなり、言語的な手がかりを用いることを特徴とする。技術論文529文書を対象とした実験を行った結果、検索語が含まれる文の情報の種類が識別された文書は、検索結果の45%から84%であった。

A METHOD OF INFORMATION RETRIEVAL BY DOCUMENT STRUCTURE ANALYSIS

Seiji Miike　Kenji Ono　Kazuo Sumita

Research and Development Center, Toshiba Corp.

Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan

This paper describes a method and its implementation of information retrieval by document structure analysis for a full-text search system. Based on extracted document structures, retrieved documents are classified on categories of information conveyed by sentences including keywords. We have implemented the method and had an experiment using 529 Japanese technical papers. As the result, from 45% to 84% retrieved documents are classified.

1. はじめに

大量の情報が利用できるようになるにつれ、情報を多様な側面、特徴から検索し利用する技術が必要になることが指摘されている[1]。また、従来の検索方法に対し、本来の情報検索を行なうためには、文章の構造に基づくことが必要であることが指摘されている[2]。このように、文書情報検索の高度化のためには、文書の論理的な構造の解析が不可欠であると考えられる。本稿で述べる検索方法は、文書構造の解析により検索語句がどのような情報・内容を述べている部分で用いられているかを識別して検索・提示するものである。筆者らは、本検索方法により、ユーザが対話的に文書情報を検索する作業を支援する全文検索システムの開発をめざしている。

キーワードの抽出で、主題文を選択しキーワードを精度高く自動抽出する研究が行なわれている([3,4]など)。筆者らは、文書中から対話的に必要な情報を探し出せるシステムを目標としており、この点がキーワード抽出の研究と異なる。また、二次情報データベースに対して、一次情報データベースが増大している[5]ことから、今後これら一次情報データベースを対象とした全文検索システムへのニーズが高まると考えられる。また、文書の主題分析や抄録作成などを目的とする文書構造分析の研究([6]など)があり、対象分野の拡張や自動化が課題となっている。筆者らは自動化の実現を図っている。

また、意味属性や主題の解析による文書のブラウジングの研究[7,8]や検索方法の提案[9]がある。これらの研究では、2文以上からなる列挙表現などの構造への対応が述べられていない。本文書構造解析方法では、言語的な手掛かりに基づいた文間の階層構造の解析[10,11,12]により、列挙表現などの構造を扱うことができる。文献[13]では、文脈の解析により文書の構造化を行なっているが、検索への応用について述べられていない。また、特定の分野や話題のニュース記事などから、そこで予想される情報(企業名や価格など)を抽出する研究([14,15]など)がある。本検索方法は、特定の分野や話題についての情報抽出に応用することが可能である。

本稿では、2節で文書構造の解析に基づく検索方法について述べ、3節で本方法による文書検索システムの文書構造解析システムについて述べる。4節で文書構造の解析に基づく検索の実験について報告し、最後にまとめを述べる。

2. 構造解析に基づく情報検索

情報検索の分野で、文書構造や文脈構造の利用が検討されている([16]など)。本構造解析では、言語的な手掛かりに基づいた解析方法を用いることを特徴とする。本研究での文書構造の解析は、文書書式の解析と、文脈構造の解析、文役割の解析の3つからなる。文書書式の解析では、タイトルや本文などの書式の階層構造を解析する。文脈構造の解析では、修辞的な接続関係("順接"や"逆接"など)に基づいて文間の階層構造を解析する。文役割の解析では、文書が伝達する情報の種類、例えば、技術論文では「目的」や、「背景」、「結論」、解説記事では「背景」や「話題」などを識別し抽出する。ここでは、文書が伝達する情報を文単位で処理し、この情報を文役割とよぶ。本検索方法では、検索語が含まれている文の文役割を区別して文書を検索し提示する。

文役割の抽出では、一文内の特定の表現を手がかりとする。従って、ある文役割をもつと判定された文が、必ずしもその文書の主題についての役割であるとは限らず、文書の主題以外の局所的な範囲での役割の場合もある。また、文書の主題についての役割であっても、その文中のすべての語句が当該の情報の種類で用いられているとは限らない。しかし、ここでは、このような場合があるとしても、全文検索の検索結果から必要な情報を得るための補助情報として有効であろうと仮定している。ユーザによる文書情報の選択を支援するために、システムは、有用と思われる情報を提示し、対話的な操作環境を提供する。上述のように、今回のインプリメントでは1文を単位として文役割を抽出するが、今後、1文内のより限定された部分から情報を抽出する方法の実現を図る。

3. 文書構造解析システム

まず、現在開発中の対話的文書検索システムの構成を図1に示す。本システムはワークステーション(東芝AS4000シリーズ)上でインプリメントされている。本システムでは、文書構造の解析によって得られる構造化データとインデックスファイルを事前に作成しておき、これを検索時に利用する方式をとっている。なお、全文検索部は中本らが開発したシステム[17]を用いている。

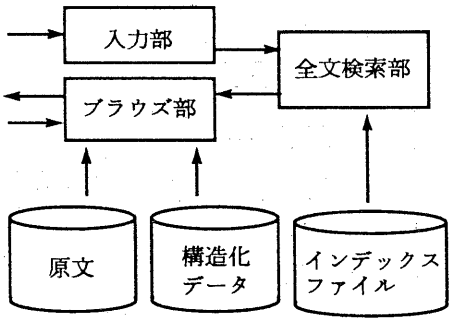


図1　文書検索システムの構成

次に、構造化データおよびインデックスファイルを作成する文書構造解析システムについて述べる。文書構造解析システムは、図2に示す7つの構成要素、すなわち文書書式解析部 、日本語解析部、文役割抽出部、文脈情報抽出部、文脈構造解析部、文役割投射部、およびインデックスファイル作成部から構成される。構造化データは、文書構造解析部、文脈構造解析部および文役割投射部の解析結果からなる。図3に解析データの例を示し、以下に図2の各構成要素について説明する。

(1) 文書書式解析部

文書書式解析部は、テキストファイル形式の文書(図3(a))から、タイトル、著者名、著者の所属先、アブストラクト、章見出しや本文などの階層構造(図3(b))を認識する。文書の自動レイアウトシステム Darwin[18]とほぼ同様の処理を行うが、現在は図表などは対象とせず、テキストのみを対象としている。見出しを認識するために、先頭が数字や記号で始まるという情報、行の末尾に句点が存在していないという情報などを文書中から取り出す。また、行の先頭が空白であることを検出し、その位置が段落の開始であることを認識する。文書書式の解析では、節ごとに見出しとその節が含む文章本体の範囲、その文章本体を構成する各段落の先頭の位置、などを構造化データとして格納する。

(2) 日本語解析部

日本語解析部では、著者名と著者の所属先以外を、見出しおよび1文ごとに形態素解析および構文意味解析する。解析辞書には約60,000語の一般用語辞書を用いている。1文全体の解析

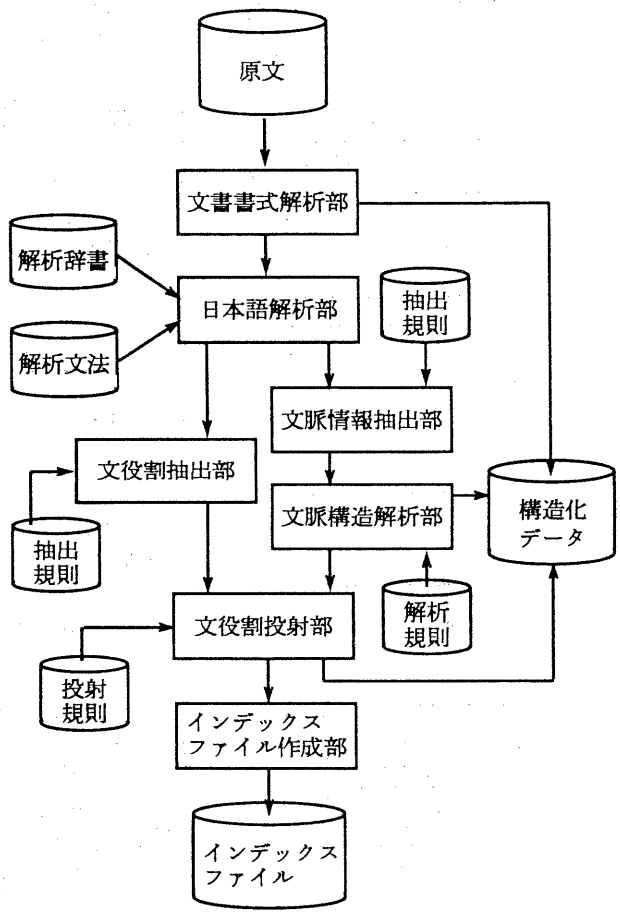
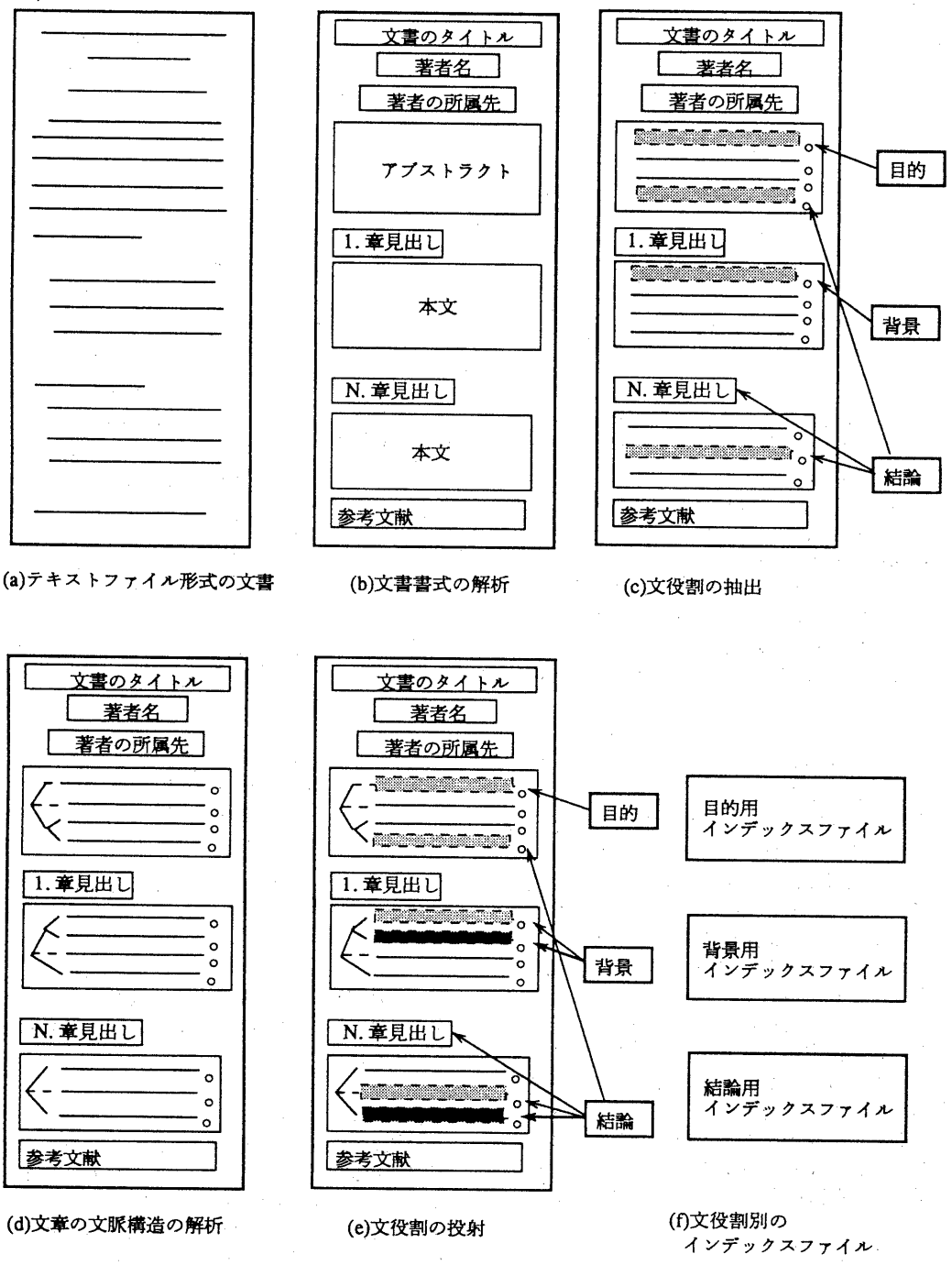


図2　文書構造解析システムの構成

木の生成に失敗した場合でも、部分の解析結果を出力する。なお、本解析部には平川らの解析システム[19,20]を用いている。

(3) 文役割抽出部

文役割抽出部は、日本語解析結果に文役割抽出規則を適用し、何文めがどの文役割であるかの情報を生成する(図3(c))。文役割抽出規則には、「近年」、「目的」のような内容語や、文末の表現、固定的な言い回しなどを手掛かりとした条件と、対応する文役割が記述されている。文役割抽出は、本文だけでなく、タイトルや章見出しにも適用する。文役割抽出部および次の文脈情報抽出部では文の一部との照合を行



(a)テキストファイル形式の文書　(b)文書書式の解析　(c)文役割の抽出

(d)文章の文脈構造の解析　(e)文役割の投射　(f)文役割別の インデックスファイル

図3　文書構造解析システムにおける解析データ

うので、日本語解析部で1文全体での解析に失敗した場合にも適用できる。

(4) 文脈情報抽出部

文書書式解析部で識別されたアブストラクトや本文などの文章部分の構造を解析するため、日本語解析結果に文脈情報抽出規則を適用して文間の接続関係を抽出する。文脈情報抽出規則には、接続詞や、文末の表現、固定的な言い回しなどの接続表現を手掛かりとした条件と、それぞれに対応する例示関係、並列関係、順接関係などの接続関係が記述されている。文脈情報抽出規則は、約1,350の接続表現と、35種類の接続関係を用いて記述されている[12]。

(5) 文脈構造解析部

文脈構造解析部は、文脈情報抽出部で抽出された接続関係と文脈構造解析規則を用いて、文間の階層構造を表現する2分木(図3(d))を生成する。文脈構造解析規則は、複数の文にわたる修辞的な表現を手掛かりとする150の規則と、接続関係の間の局所的な構造に関する600のプリファレンス規則からなる[12]。

(6) 文役割投射部

文役割投射部では、文役割抽出部で抽出された文役割の情報を、文脈構造解析部で生成された文間の階層構造に照合し、文役割投射規則に従って文役割を複写する(図3(e))。例えば、「以下の特徴がある。....。....。」のような列挙表現では、文役割抽出部で1文目から文役割「特徴」が抽出され、文脈構造解析部で列挙表現と認識された範囲の文に、この文役割を付与する。

(7) インデックスファイル作成部

テキストデータベースの各文書について以上の処理を行ない、文役割ごとのインデックスファイル(図3(f))を作成する。インデックスファイルは、各文役割について、その文役割を担っていると認識された文から抽出された単語と、全文書の文書IDからなるテーブルである[17]。

3.2 構造解析の例

文脈構造解析による文役割投射の例を次に示す。この文書では、下記の「まえがき」の中の1文のみに「インタフェース」が含まれている。なお、文頭の(1)と(2)および下線は筆者が挿入したものである。この文書では、文(1)の「近年、ーている。」が文役割抽出規則の条件と一致し、文役割「背景」が付与された。一方、文脈構造解析によって、文(2)の文頭の「それに伴い」という表現から、文(1)と(2)の接続関係が順接関係とされた。さらに、文役割投射部において、文(1)の「背景」が文(2)に複写された。その結果、「インタフェース」が「背景」の文で用いられていることが識別された。

文書の例："AIワークステーション",斎藤他,
東芝レビュー,Vol.42,No.5(部分)

〔1〕まえがき

(1)近年、人工知能の研究が急速に実用化に向けて進み、さまざまな応用システムの開発が進められている。(2)それに伴い、応用システムを効率良く開発し、かつ実用的な速度で実行するAI(Artifitial Intelligence)用計算機、特にマンマシンインタフェースのよいAIワークステーションの開発が強く望まれている。

従来のAIワークステーションは、LISPまたはPROLOGなどのAI言語専用ワークステーションか、または通常のEWS(エンジニアリングワークステーション)に単にAI言語処理系を載せたものであった。そのため、前者においては,通常の言語で書かれた既存のプログラムを実行することができず、後者においては、AI用言語で書かれたプログラムの実行速度が十分でなかった。

今回開発したAIワークステーションは、従来のEWS(AS3000シリーズ)をベースに、AIプロセッサ(以下AIPと略称する)を開発し付加することにより、両者の機能を兼ね備えた実用的なものにするもので、広い応用範囲が期待されるものである。ここでは、そのアーキテクチャの設計方針と、その評価について述べる。

4.検索実験

技術論文を対象とし、文書構造の解析を行わない場合の検索結果と、文書構造の解析を行う場合の検索結果との比較調査を目的とした実験を行った。ここでは、文役割の抽出の頻度を中心として述べる。文中のどれだけの単語が、文役割として抽出された情報の種類(「目的」や「結論」)で用いられているかなどの評価については稿を改めて報告する。

4.1 実験の方法

検索対象に技術論文誌「東芝レビュー」を用いた。東芝レビューを調査し、文役割の種類として表1に示す6種類を用意し、文役割抽出規則を122記述した。表1に、文役割抽出の手掛かりとしていた表現の例も示す。

表1 文役割の種類、抽出規則の数および表現の例

種類	規則数	表現の例
話題	15	ーについて解説する
目的	21	ーを目的としたものである
背景	41	ーしてきた
特徴	11	(ーの)特徴はーである
結論	22	ーを図った
課題	12	ーをめざしたい
合計	122	

表1の「特徴」には、利点や長所などに関する表現も含め、「結論」には、結果など事実を現す表現も含めている。文役割抽出規則以外の辞書や規則には、3節で述べたものを使用した。文書構造解析システムにより、529文書(6.1Mバイト)の構造解析を行った。

4.2 実験結果

4.2.1 文役割の抽出頻度

14文書について、文書書式と文の数を調べた。表2(A)に、(i)アブストラクト、(ii)最初の章、(iii)最初の章と最後の章以外の章、および(iv)最後の章について、その中に含まれる文の数の平均値を示す。表5(B)と(C)は、それぞれ文役割が抽出された文の数の平均値とその比率を示す。アブストラクトと最初の章、最後の章からは、5割前後の文から、またそれ以外の章からは2割の文から文役割が抽出された。また、表3に、上記の(i)から(iV)について、文役割ごとにその中の文から文役割が抽出された文書の数を示す。

4.2.2 検索結果の文書数

(1)「インタフェース」、(2)「計算機」、(3)「原子力発電所」、(4)「設計」の4単語を用いて検索を行った。各検索語について、検索された文書の数を表4に示す。「検索語」の列の(1)から(4)は、上記の検索語に対応する。(A)の列は、構造化を行わない場合の文書数である。(B)の列は、検索語がいずれかの文役割の文に含まれていた文書の数である。(C)の列は、(A)に対する(B)の比率(B/A)である。また、その下に文役割別にその文役割の文に検索語を含んでいた文書の数を示している。表4に示すように、構造化を行わない場合に検索される文書の45%から84%の文書について、検索語が含まれる文の文役割が識別された。

次に「背景」と「結論」およびそれら以外の文役割を「その他」として、各検索語を含む文書の分布を図4に示す。図4中の数字は文書の数を表している。なお、3.2の説明に用いた文書は、図4(1)で検索語が「背景」の文でのみ使われている7文書の中の一つである。

4.2.3 検索結果の絞り込みの比率

例えば、次の4つの順序に従って、文書を検索する方法が考えられる。次の(a)から(d)は図5中のa, b, c, dに対応する。

(a) 検索語を結論の文に含む
(b) 検索語を「その他」の文役割の文に含む
(c) 文役割が抽出されない
(d) 検索語を背景の文のみに含む

図4に示した結果について、上記の絞り込みの率を算出した結果を表3に示す。表5において、(1)から(4)は検索語の番号に対応し、(a)から(d)は上記の分類に対応する。また、各検索語についての値では、左の列が上記分類に対応

表2 文書書式と文役割

	(i)	(ii)	(iii)	(iv)	全体
(A)	4.2	6.2	62.2	3.8	76.2
(B)	2.0	2.5	12.0	2.1	18.6
(C)	47.6%	40.3%	19.3%	55.3%	24.7%

表3 文書書式と文役割

	(i)	(ii)	(iii)	(iv)
話題	5	7	7	3
目的	2	2	11	1
背景	4	10	6	2
特徴	0	1	3	0
結論	5	8	9	12
課題	0	1	1	4

表4 検索語を含む文の役割別の文書数

検索語	(1)	(2)	(3)	(4)
(A)	156	138	37	319
(B)	70	86	31	220
(C)	45%	62%	84%	69%
(内訳)				
話題	37	47	17	137
目的	9	15	7	56
背景	18	32	15	69
特徴	3	4	2	13
結論	37	46	22	148
課題	2	2	0	8
合計	106	146	63	431
異なり	70	86	31	220

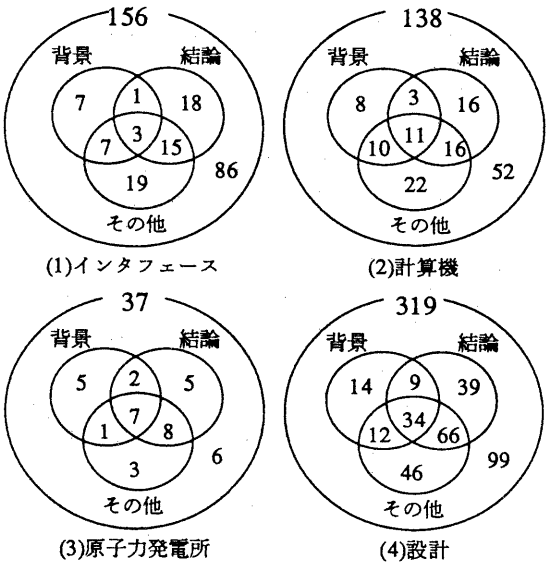


図4　検索語を「背景」、「結論」または「その他」の文に含む文書の数

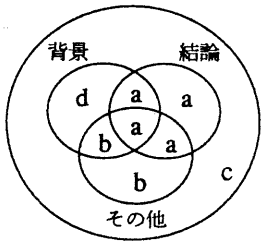


図5　検索文書の分類

表5　検索文書の絞り込み率

	(1)		(2)		(3)		(4)	
(a)	37	24	46	33	22	59	148	46
(b)	26	40	32	57	4	70	58	65
(c)	86	96	52	94	6	86	99	96
(d)	7	100	8	100	5	100	14	100
合計	156		138		37		319	

表6　文役割が識別された文書の比率

	(1)		(2)		(3)		(4)	
(a)	37	53	46	53	22	71	148	67
(b)	26	37	32	37	4	13	58	26
(d)	7	10	8	9	5	16	14	6
合計	70	100	86	100	31	100	220	100

する文書の数であり、右の列が(a)から(d)の順序で累積した文書の絞り込み率(%)である。また、文役割が識別された文書について、(a), (b)および(c)の割合を表6に示す。(1)から(4)および(a), (b), (d)の意味は表5と同様である。各検索語の左の列は同様に分類に対応する文書の数であるが、右の列は文役割が識別された文書の中での割合(%)を示す。

表5から、検索語を含む文書全体に対して、検索語を文役割が結論である文に含む文書の割合は、24%から59%であった。同様に、検索語を背景以外の文に含む文書の割合は、40%から70%であった。また、背景の文のみに含む文書の割合は、4%から14%であった。表6から、文役割が抽出された文書の53%から71%が、文役割が結論である文を含んでいた。

5. おわりに

文書の論理的な構造を解析する方法と、その解析結果を利用する検索システムについて述べた。実験の結果、検索語がどのような情報を伝える文の中に存在するかを、45%から84%で区別した。ユーザに、検索語が「結論」の文に含まれる文書であるか、「背景」の文に含まれる文書であるかなどを区別して提示することは、検索作業の効率化のために有用であると考えられる。本実験から、その支援機能の実現の見通しを得た。

また、文書を構造化することにより、次のような検索支援機能の実現が考えられる。

・検索命令の入力において、例えば「目的」に「画像」と「処理」を含み、「結論」に「速度」を含むというような指定を可能にする。また、入力命令の解析を行うことにより、「画像処理を目的とし、速度に関する結果について述べた文書」のようなよりユーザにとってより自然な対話入力を実現する。

・ユーザは文役割などの指定をせず、システムが文書の構造解析結果を用いて、検索語がどのような文役割で用いられている文書がいくつあるかを提示する。複数の検索語が用いられた場合には、さらに、すべての検索語が当該の文役割の文に含まれる文書から、すべての検索語が当該の文役割でない文に含まれる文書まで分類して表示する。このようにユーザが多面的に文

書群を見ることを可能にし、ユーザの検索戦略の立案を支援する。

今後、文書の構造化において、1文内の部分から「目的」や「結論」などに該当する情報を取り出す機能、および単語間の格関係を抽出する機能などを実現し、検索結果の精度向上を図る。

[謝辞]
本稿で述べた文書検索システムの研究開発に際して、ご指導や討論をして頂いた東芝情報処理・機器技術研究所の岩井勇氏、中本幸夫氏と、東芝研究開発センターの竹林洋一、平川秀樹、武田公人、伊藤悦雄、水谷由美、田中克己の諸氏に感謝する。

[参考文献]
[1]長尾真、情報社会の生態学、情処学会人文科学とコンピュータ研究会 11-6, 1991.
[2]細野公男、情報検索理論・技法の問題点とその解決の方向、情処学会情報学基礎研究会 FI-24-5, 1991.
[3]石川徹也、文意解析処理に基づく主題索引語作成支援システム、情報処理学会論文誌 Vol. 32, No.2, 1991.
[4]原田隆史他、主題文を対象とした索引語の自動抽出、情処学会情報学基礎研究会FI-29-3, 1993.
[5]三輪真木子、データベースサーチャの視点、情報処理 Vol.33, No.10, 1992.
[6]神門典子、構成要素カテゴリを用いた原著論文の内部構造分析、情処学会情報学基礎研究会 FI-25-7, 1992.
[7]中本幸夫他、文書への意味属性付与のための意味辞書の拡張、情処学会第45回全国大会 6F-7, pp.3-211-212, 1992.
[8]野上謙一他、文書の意味構造を用いたブラウジング機能の開発、情処学会第45回全国大会 5F-8, pp.3-193-194, 1992.
[9]新開正史他、主題の意味構造に基づく論文検索法の提案、情処学会第46回全国大会 6G-6, pp. 4-207-208, 1993.
[10]小野顕司他、文脈構造の分析、情処学会自然言語処理研究会NL-70-2, 1989.
[11]Sumita, K., et al., A Discourse Structure Analyzer for Japanese Text, Proc. Int. Conf. Fifth Generation Computer Systems 1992(FGCS'92), pp. 1133-1140, 1992.
[12]小野顕司他、日本語論説文の自動抄録のための文脈構造解析、情処学会第46回全国大会 7B-10, pp.3-187-188, 1993.
[13]西村健士他、特定表現の重点解析による科学技術論文構造化手法、情処学会情報学基礎研究会 FI-29-5, 1993.
[14]Jacobs, P.S., et al. SCISOR: Extracting Information from On-line News. CACM, Vol. 33, No. 11, pp.88-97, 1990.
[15]須之内美幸他、フルテキストの構造化に基づく検索システム、情処学会データベースシステム研究会 DBS-90-7, 1992.
[16]Hearst, M.A., Cases As Structured Indices for Full-Length Documents, AIII'93, Spring symposium series, CBR/IR: Exploring Opportunities for Technology Sharing, pp.140-145, 1993.
[17]中本幸夫他、日本語解析を用いたフルテキストサーチの実験、情処学会第46回全国大会 4B-4, pp.3-125-126, 1993.
[18]Iwai, I., et al. A Document Layout System Using Automatic Document Architecture Extraction. Proc of CHI'89, pp.369-374, 1989.
[19]平川秀樹他、構文/意味優先規則による日本語解析、人工知能学会第3回全国大会 8-4、1989.
[20]平川秀樹他、日本語解析における最適解探索、情処学会自然言語処理研究会 NL-74、1989.
[21]原田隆史他、構文解析に基づくキーワードへのロール決定方法の高度化、第27回情報科学技術研究集会, 1990.
[22]住田一男他、対話的文書検索のための文書構造解析、情処学会自然言語処理研究会 NL-97、1993.