自動抄録機能をもつ対話的文書検索システム
2V－8　— 検索機能 —

三池誠司　小野顕司　住田一男
(株)東芝　研究開発センター

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、文章の内容に基づくことが必要であることが指摘されている[2]。

我々は、効率的な検索を目的とし、文書の構造に基づく検索を行う文書検索システムBREVIDOCを試作した[3]。本稿では、本システムの検索機能について述べる。

2. 文の意味役割に基づく検索

検索の高度化のため文書構造や文脈構造の解析を行う方法が検討されている([4]など)。本稿で述べる検索方法では、全文検索をベースとし、検索語句がどのような情報・内容を述べている文で用いられているかを解析し検索することを特徴とする。この情報を文の意味役割とよぶ。文の意味役割は、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である。この解析を行うために、文書構造解析システム[5]を開発した。このモジュールは検索に先立って起動される。

3. 文書構造解析システム

図1に文書構造解析システムの構成を示す。文書構造解析は、破線で図示した文書構造解析[6]と実線で図示したインデックス作成の処理からなる。構造化データは、抄録生成処理において参照される[7]。ここでは、インデックス作成処理について述べる。

意味役割抽出部は、日本語解析結果に文役割抽出規則を適用し、何文目がどの意味役割であるかの情報を生成する。意味役割抽出規則には、「近年」、「目的」のような内容語や文末の表現などを手掛かりとした条件と、対応する意味役割が記述される。

意味役割投射部では、意味役割抽出部で抽出された意味役割の情報を、修辞構造解析部で生成された文間の階層構造と照合し、意味役割を複写する。例えば、「以下の特徴がある。．．．．。．．．．。」のような列挙表現の場合、文役割抽出部で1文目から文役割「特徴」が抽出

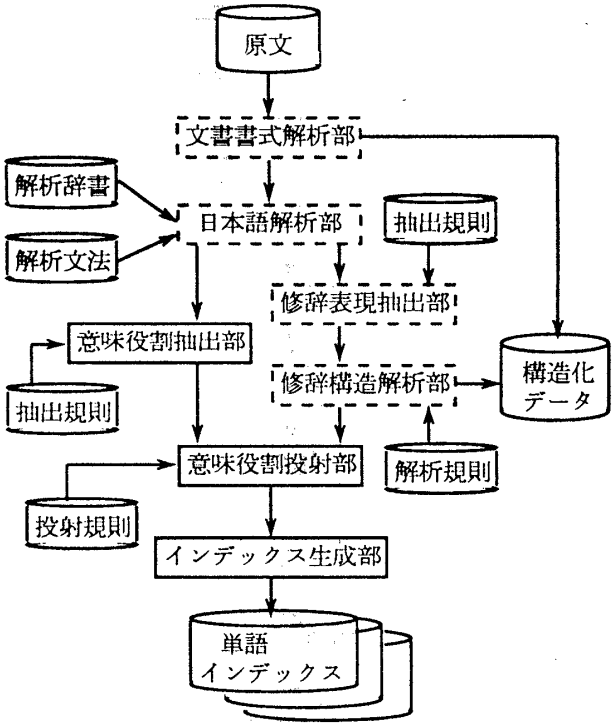

図1　文書構造解析システムの構成

される。修辞構造解析部で列挙表現の範囲が認識されるので、それらの文に意味役割「特徴」を付与する。

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、意味役割ごとのインデックスを作成する。インデックスは、各意味役割について、その意味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[8]である。

4. 検索の例

検索部は、「目的:被爆/低減」のような検索コマンドに従い、その中の意味役割、この場合は「目的」に対応する単語インデックスを参照し文書を検索する。この検索コマンドで、技術論文誌である「東芝レビュー」386文書を検索した結果、次の文書を含む10文書が得られた。

可児次郎他：「東北電力(株)女川原子力発電所第1号機の建設」, Vol.39, No.11, 1984.

上記の文書は3章に次の文を含んでいる。この文は下線で示したように「ーを目的とする」の表現を含むことから、文書構造解析システムによりこの文に含まれる単語が「目的」の意味役割のインデックスに登録され、当

A Document Retrieval System with an Interactive Abstract Generator - Retrieval Function -
Seiji MIIKE, Kenji ONO and Kazuo SUMITA
Research and Development Center, Toshiba Corp.
miike@isl.rdc.toshiba.co.jp

該の文書が検索されている。

クリーンプラント作戦では放射性廃棄物発生量の低減のためのクラッド低減および系統機器配管内面の清浄度維持による炉内へのクラッド流入の抑制および運開後の被爆低減を目的とするもので，図4に示す具体的な施策を計画的に実施した。

5. 評価

5.1 文の意味役割抽出

「東芝レビュー」386文書を対象とし、話題、目的、背景、特徴、結論、課題の6種類の意味役割について、122の意味役割抽出規則を記述し文書構造解析を行った。その中の14文書の1,071文について意味役割の抽出精度を評価した。意味役割は261文から抽出され、これらの中で253文は正しく抽出された。意味役割をもつ文でそれが抽出されなかった文は62文であった。従って、意味役割の適合率は96%であり、再現率は80%であった。再現率が低かったことの主な原因は、日本語解析処理または修辞表現抽出・修辞構造解析処理の失敗、および意味役割抽出規則の欠如であった。

検索性能の評価のため、「被爆/低減」、「CCD/イメージセンサ」、「ソフトウェア/再利用」を含む各10文書、10文書、13文書について、それらの文書から抽出された話題、目的、背景、結論の意味役割の適合率と再現率を調べた。結果を表1に示す。表1の値は上記の3組の単語を用いた検索の平均値である。

表1 文の意味役割の適合率と再現率の平均値

	話題	目的	背景	結論
適合率	100	90	42	65
再現率	57	62	19	71

特に話題と目的では適合率が高かった。

意味役割の抽出に関しては、さらに実験評価を行い、抽出精度、特に再現率を高めていく必要がある。

5.2 本検索機能の有効性

文書検索の目的の一つは、利用者の検索要求に適合する、すなわちその内容が中心的に述べられている文書（適合文書）を探すことである。その観点から、本方式の有効性を確認することを目的とし、全文検索方式と比較評価した。文書の中に中心的に書かれる内容と関係が深いと考えられる意味役割は、上記の6種類の意味役割の中では、話題、目的、特徴および結論であるので、ここでは、検索コマンドに設定する文の意味役割をこの4種類の意味役割とした。

表2に結果を示す。表2の左側の単語は、検索コマンドに含まれた単語(検索語)であり、AからGの欄は次の値を示す。

A:全文検索により得られた文書の数
B:人手により判別したAの文書中の適合文書の数
C:全文検索の場合の適合率(B/Aの値(%))
D:検索語が話題、目的、特徴または結論に含まれた文書の数
E:Dの文書の中に含まれた適合文書の数
F:本方式の場合の適合率(E/Dの値(%))
G:適合文書の比較(E/Bの値(%))

表2 適合率の評価結果

	A	B	C	D	E	F	G
1 被爆/低減	10	6	60	7	6	86	100
2 CCD/イメージセンサ	10	5	50	7	5	71	100
3 ソフトウェア/再利用	13	3	23	4	3	75	100
4 保守/エキスパートシステム	13	3	23	5	3	75	100
5 原子力発電所/設計	20	3	15	12	2	17	67
6 映像/信号	26	12	46	7	6	86	50
平均			36			66	86

全文検索の適合率が平均36%であるのに対し、本方式では平均66%の値を得た。ただし、表2の5番目と6番目のように、適合率に差がない場合(5番目のCとFの値)や、本方式で適合した文書の数が全文検索方式のそれを大きく下まわる場合(Gの値)があった。

本方式は、再現率が低いことがあるが適合率が高いことから、レレバンス・フィードバックのように対話的に検索精度を高めていく機能には整合性がよいと考えられる。

6. おわりに

試作した自動抄録機能をもつ文書検索システムの検索機能について述べた。本検索機能は、全文検索をベースとし、検索語句が含まれる文の意味役割を識別して検索する。意味役割の抽出は文を単位とするが、利用者の検索要求に適合する文書を検索する観点から有効であることを確認した。今後、意味役割の抽出精度向上と、文書情報の可視化提示、レレバンス・フィードバックへの応用を図る予定である。

参考文献

[1] 長尾真：「情報社会の生態学」、情処研資 CH-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」、情処研資 FI-24-5, 1991.
[3] 住田一男他：「自動抄録機能をもつ対話的文書検索システムーシステムの構成と機能ー」、第48回情全大, 第3分冊, 2V-7, 1994.
[4] Liddy, E.D., *et al* : "DR-LINK Project Description", *SIGIR Forum*, Vol.26, No.2, pp.39-43, 1992.
[5] 三池誠司他:「文書の構造解析に基づく文書情報検索」、情処研資 FI-31-6, 1993.
[6] Sumita, K., *et al* : "Document Structure Extraction for Interactive Document Retrieval Systems", *Proc. SIGDOC'93*, pp.301-310, 1993.
[7] 小野顕司他：「自動抄録機能をもつ対話的文書検索システムー自動抄録機能ー」、第48回情全大, 第3分冊, 2V-9, 1994.
[8] 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」、第46回情全大, 第3分冊, 4B-4, 1993.