データベースシステム 90－7
（1992. 9. 11）

フルテキストの構造化に基づく検索システム

須之内美幸[1)]、岸本行生[1)]、塚田康博[1)]、千葉滋[1)]
石川徹也[2)]

[1)]シャープ株式会社
[2)]図書館情報大学

（抄録）

大容量のテキストを対象とする検索システムには、多様な検索要求に対し適合率の高いシステム機能が求められる。その為には、従来の主題表示索引語に対し、検索要求意図を忠実に反映するフルテキスト対応の検索システムが必要になる。そこで、筆者らは、上記システム機能の実現の為に、フルテキストの日本語ニュースと質問文の内容をそれぞれの意味構造に変換し、それらの構造間の比較・照合によって適合するテキストを検索するシステムの開発を計った。本稿では、システム機能を中心に紹介する。

An Information Retrieval system Based on Full-text Semantic Analysis

Miyuki Sunouchi[1)], Yukio Kishimoto[1)], Yasuhiro Tsukada[1)], Shigeru Chiba[1)]
Tetsuya Ishikawa[2)]

[1)]Sharp Corporation
[2)]University of Library & Information Society

Abstract

Information retrieval systems intended for use with large text databases necessitate system functions highly compatible with diverse retrieval requests. For this, it is necessary that the retrieval system, in opposition to the Subject indexing terms method, correspond to full-text data and faithfully reflect a user's retrieval intention. To realize the above stated functions, we have planned the development of a system that takes the content of both full-text Japanese language news and also the user's request and alters each into a semantic structure, and then retrieves text by performing comparative texts among these structures. In this paper, we principally introduce these system functions.

1.はじめに

テキスト・データベース（以下、テキストDBと記す）を対象とする検索システムには、多様な検索要求に対し適合率の高いシステム機能が求められる。その為には、検索要求意図を忠実に反映するフルテキスト対応の検索システムが必要になる。筆者らは、上記システム機能の実現の為に、フル・テキストの日本語ニュースと質問文を意味構造に変換し、質問文の意図を忠実に反映し検索するシステムの開発を計った。本稿では、当システムの機能を中心に紹介する。

以下、2.でテキスト・データベースを対象とする現行の検索システムの課題について述べ、3.で筆者らの開発しているシステム機能を紹介し、4.で当システム機能の有効性と今後の課題について述べる。

2.テキスト検索システムの課題

2.1　現行のテキスト検索システムの問題点

文書、論文等テキストを検索する場合、書誌データ（例：著者名、表題、出版社名等）を対象に検索する場合と、テキストの内容を対象に検索する場合とがある[注]。テキストの内容を対象に検索することを可能にする為に、図1に示すように、現行のシステムでは、検索用データとしてテキスト内容を代表する索引語が用いられている。

索引語は、従来から、索引語作成者がテキストの内容を解読し、シソーラス用語（統制索引語）を用い設定してきている（当方式を事前索引方式(Pre-indexing method)と呼ぶ）。しかし、大量のテキストに対し、索引語作成者による索引語作成には限界があり、情報提供に対するタイム・ラグ、索引語の質の揺れ等の問題点が生じ、検索精度に影響を及ぼしている。当問題回避の為に、テキストの内容を自動解析し、テキスト内に出現する単語を対象にキーワードとして抽出し、索引語とするキーワード自動抽出システムが利用されるようになってきている。(2)

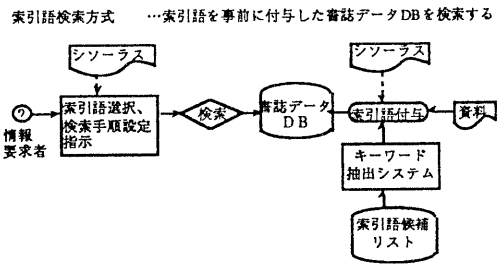


図1.　索引語検索方式のテキストＤＢシステム

現在、実用に供されているキーワード自動抽出システムを大別すると、次の２方式のものがある。

（１）事前設定による大規模索引語候補語リストを用い、テキスト内に出現する単語との照合によりキーワードを自動抽出するシステム。例：(1),(3),(4)

（２）検索語に対し、テキスト内に出現する単語をキーワードとして自動抽出するシステム。例：(5)

また、キーワードによる検索システムを含めフルテキスト検索方式には、図２に示すような２種類のシステムが挙げられている(6)。

しかし、いずれのシステムもテキスト内に出現する文字列との照合を前提として、意味に関係なく抽出する為、再現率は補償されるが、適合率は補償されないという問題がある。適合率を補償する為には、質問文およびテキストを解析し、検索要求の意図を忠実に反映した検索を行なう必要がある(6)。

[注]　テキストを対象とするDBとして、下記の２種類が構築、提供されている。ひとつは書誌データと索引語（以下、両者を検索用データと呼ぶ）を蓄積した書誌DBであり、もうひとつはフルテキストDBである。このことに対して、検索用データの設定方式に、下記の２方式がある。ひとつは検索用データをテキストの内容解析の基に事前に設定しておく方式（事前設定方式と呼ぶ）であり、もうひとつは検索指示の都度検索指示内容に適合する検索用データをテキスト内に発見し検索する方式（フルテキスト検索方式と呼ぶ）である。

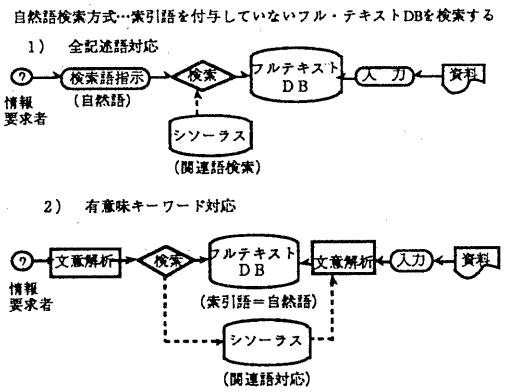


図2 自然語検索方式のテキストDBシステム

2.2.テキスト検索システム機能の高度化

情報検索要求には、例えば「Aについて知りたい」という情報要求に対して、検索結果に対する期待として次の2種類があるものと考える。

一つは、検索要求事項「A」に"関連する情報"の検索結果を期待する場合と、もう一つは、例えば「Aを購入したいので」と言った"隠された願望・目的"と「但し、Aは、1,000円以下」といった"検索条件"に見合う検索結果を期待する場合である。前者を関連情報検索と呼び、後者を知識情報検索と呼ぶことができる(7),(8)。しかし、現行のキーワード自動抽出システムでは、上記の関連情報検索および知識情報検索に対応するキーワードは抽出できない。

これに対して、関連情報検索システムとして、システム内にシソーラスを取り込み、指示検索語とシソーラス用語とを先ず照合し、該当するシソーラス用語を含むカテゴリ内の全ての用語を基にフルテキスト内に出現する単語に対し検索を行なうシステムが実用化されている(9)。しかし、当システムもテキスト内に出現する全ての単語を対象に検索を行なう為、検索精度に対し補償できないという欠点があり、さらに、シソーラスを常にメンテナンスしなければならないという問題がある。

知識情報検索については、例えば、「購入したい」という"隠された願望・目的"に対し、「既に販売されている」という検索結果でなければ意味をなさないし、「A」が「1,000円以下」の検索結果でなければ意味をなさない。これに対し、現行のキーワード抽出システムでは、検索要求を満たすことはできず、「全てのA」に関するテキストを検索結果として、後は検索利用者の判断に任せる方式になっている。

上記の問題に対しては、例えば、"検索条件"に見合う検索用データ項目として、「価格」等のファセット化あるいはオブジェクト化が提案されている(10)。しかし、誰もが共通に認識できる普遍的ファセットあるいはオブジェクト項目の設定は、実際には困難であり、検索精度に問題を残している。ましてや、"隠された願望・目的"に対応するシステム化は行なわれていない。検索の適合率を上げるには、上記の"隠された願望・目的"および"検索条件"に対応する検索機能の実現を計る必要がある。その為には、質問文およびテキストを解析して構造化し、検索要求意図を忠実に反映して検索するシステム化を計る必要がある。

3.システム機能

3.1 システム機能の概要

上記問題点を踏まえ、筆者らは、フルテキストの日本語ニュースを対象に、自然言語入力の質問文で検索するシステムを開発している。本システムは、対象テキストと質問文の内容をそれぞれ意味構造に変換し、それらの比較・照合によって適合するテキストを検索する。図3にシステム機能の概要を示す。

3.1.1 検索要求の種類

本システムは、ユーザの多様な検索要求に応えることを目標にしている。検索要求を大別すると以下の2つが考えられる。

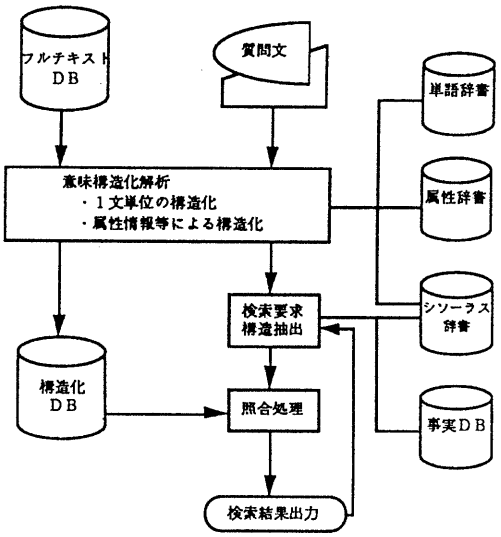


図3　システム機能概要図

（1）関連情報検索

質問文の内容と同等の情報を含む記事のみを検索する。具体的には次の2つの機能要求がある。

・関連情報検索

一文に表れる関連した文を含む記事を検索する。

・文脈対応検索

複数の文にまたがって表れる関連した文を含む記事を検索する。

（2）知識情報検索

質問文の内容に対して、知識を補うことで、ユーザの隠れた願望や目的等の意図を含む情報を検索する。具体的には、次の4つの検索要求がある。

・意図反映検索

あいまいな検索要求から検索意図を抽出し、検索要求意図を反映した記事を検索する。

・範囲対応検索

数値に対して範囲指定に応じた記事を検索する。

・事実ベース検索

検索ユーザの持っている知識を利用して記事を検索する。

・連続検索支援

連続した検索を効率良く行なう。

各々の機能の内容については次節で説明する。

3.1.2 テキストデータの構造化

上記2種類の検索要求を満たす為には、対象テキスト及び質問文の内容を忠実に理解し、テキストの文章間の関連性を理解する必要がある。この必要性を満たす為に、筆者らは、質問文及び全対象テキストに対して意味構造解析を行い、構造化する手法を提案する。その意味構造は、図4のような構造になる。

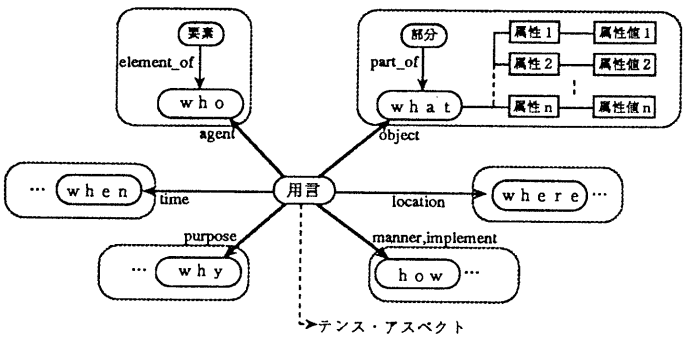


図4　意味構造モデル

意味構造への構造化は、次の２つのステップで行なう。

（１）一文単位の構造化

質問文及びテキストを一文毎に意味解析し、動詞を中心にした５Ｗ１Ｈの関係構造表現に変換する。テンス、アスペクト情報は動詞の修飾情報として扱う。図５に１文単位の構造化例を示す。

例文１：米IBM社は３月に日本でパソコンを発売した。

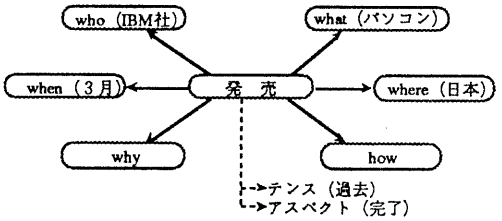


図５　一文単位の構造化例

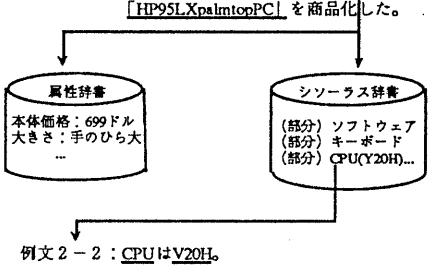
（２）名詞句のリンクによる文章間の接続

上記の一文単位の構造における５Ｗ１Ｈに相当する名詞句に対して、属性情報やシソーラスの全体／部分情報を利用して、文章中の名詞句間をリンクし、文章全体を構造化する。属性情報とは、名詞の特徴を示し、（対象ー属性名ー属性値）の組み合わせで表現する。シソーラスの全体／部分情報は、名詞間の構成要素関係を示す。ここで用いる属性情報、全体／部分情報は、それぞれ属性辞書、シソーラス辞書の一部としてあらかじめ用意する。

名詞句のリンクによる、文章間の接続例を図６に示す。

まず、動詞「商品化する」の対象格に相当する「パソコン」に対して、上記の属性辞書およびシソーラス辞書から属性情報と全体／部分情報を抽出する。次に、その情報に該当する名詞句を同一文中及び以降の文から探す。名詞句の情報を比較し、同一であればリンク付けを行ない、文章間の接続を行なう。

例文２－１：HP社とLotus社は共同でIBMPCXT互換パソコン「HP95LXpalmtopPC」を商品化した。



例文２－２：CPUはV20H。

図６　名詞句リンクによる文章間接続例

対象テキストに対しては、あらかじめこの２段階の構造化を行なってデータベース化しておく。質問文に対しては、入力毎に解析を行なう。質問文とテキストの照合は、構造間のマッチングによって行なうが、その際にこのリンクを利用する。

3.2 検索システム機能

3.1.1で挙げた検索要求に対応したシステム機能を質問文及び検索結果の例を示しながら、以下に説明する。

3.2.1 関連情報検索

（１）関連情報検索機能

自然言語で入力した質問文に対して、関連情報を含むフルテキストの日本語ニュースを検索する機能である。ここで関連情報とは、「動詞を中心にした５Ｗ１Ｈの構造やその名詞間の関係による構造が同等である情報」である。また、名詞句に関する情報とは「シソーラスの上位／下位及び全体／部分関係と属性情報による関係」である。以下に本機能の質問例とその検索結果例を示す。

［質問文１］

インテルが開発したコンピューターについて知りたい。

［検索結果１］

米Intel Corp.が並列スーパーコンピューターを開発した。

●システム機能

質問文及びテキストの内容を解析し、それぞれ動詞を中心にした意味構造を抽出する。［質問文１］の意味構造は図７のようになる。

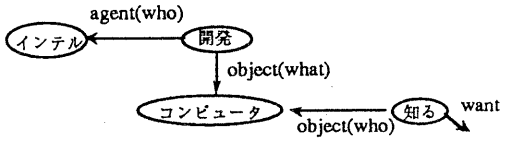


図７　意味構造例１

なお、上記意味構造中のノードは表層の単語ではなく、意味を表す。また、本論中においても、表層の単語は「コンピューター」で表し、意味は＜コンピューター＞で表す。

この意味構造から図８の網掛の検索要求構造部分を抽出する。

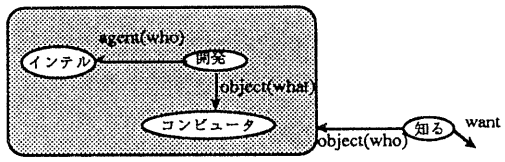


図８　検索要求構造

上記［質問文１］の意味構造中の＜コンピューター＞というノードに対してシソーラス展開を行い、図９に示す［検索結果１］の構造と照合する。

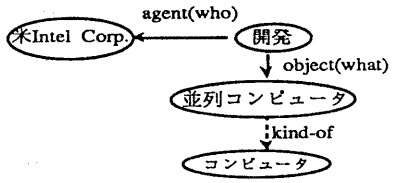


図９　意味構造例２

上記のような構造と照合することで、キーワード検索方式に見られるような、表層の単語のみが該当するニュース記事は検索しない。以下に非検索例を示す。

［非検索例１］

沖電気工業は、米Intel Corp.のNIIを搭載したミニコンを開発した。

［非検索例１］では、［質問文１］と同じ「インテル（Intel Corp.）」、「開発」、「コンピュータ（ミニコン）」といった単語を含んでいる。しかし、そこでは「インテル」という単語の意味構造における意味的な役割が異なるので、検索されない。図１０に［非検索例１］の意味構造を示す。

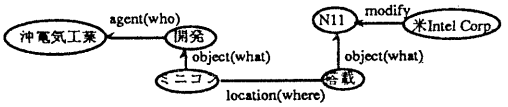


図１０　意味構造例３

（２）文脈対応検索機能

対象テキストの主題ではなく、その主題に関して述べられている情報を含む質問文で該当テキストを検索することができる。

以下に本機能の質問例とその検索結果例を示す。

［質問文２］

V20Hを搭載するコンピュータについて知りたい。

［検索結果２］

HP社とLotus社は共同でIBM PC互換パソコンを商品化した。・・・＜文１＞

…

CPUはV20H。・・・　＜文３＞

●システム機能

複数の文にまたがって記述されている内容でも、名詞句に対してシソーラスの上位／下位関係、全体部分関係と属性情報による関係を利用してあらかじめ構造化しておくことで検索を可能とする。

［質問文２］の＜V20H＞は、コンピュータ

のCPUである。一方、［検索結果２］では、＜文１＞と＜文３＞で、同様の内容を記述している。従来、単文単位で処理する検索システムでは、複数の文にまたがって記述されている内容を検索することはできなかった。しかし、上記＜文１＞の＜パソコン＞と＜文３＞の＜CPU＞には、コンピュータとその構成要素という意味的な関係がある。その関係を用いてあらかじめ名詞句間を文間接続する為、文間を越えた、検索が可能となる。

上記の［質問文２］の意味構造を図１１に示す。

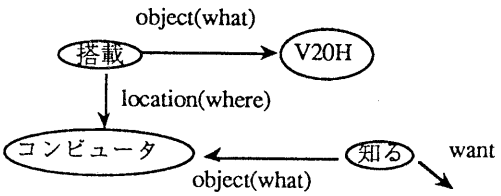


図１１　意味構造例３

また、［検索結果２］の意味構造を図１２に示す。

この例では、＜搭載する＞ということと＜パソコン＞と＜CPU＞の間の「全体／部分関係」が同等であることも利用している。

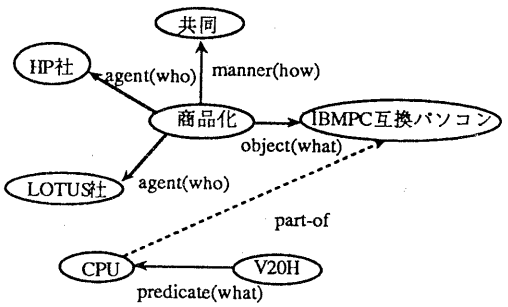


図１２　意味構造例４

3.2 2 知識情報検索

あいまいな検索要求に対して情報を補い、検索要求意図を反映して検索する機能である。具体的には以下の3つの機能がある。

(1) 意図反映検索

あいまいな検索要求から検索意図を抽出し、検索要求意図を反映した検索を行う機能である。以下に本機能の質問例とその検索結果例を示す。

［質問文３］

V20Hを搭載するパソコンを購入したい。

［検索結果３］

V20Hを搭載するパソコンが販売された。

●システム機能

上記のように、「購入する」という検索要求に応える為には「パソコンが販売されている」という前提が必要である。このような前提となる関係に対応する為に動詞シソーラスを利用して、動詞の展開を行う。ここで用いる動詞シソーラスには、検索要求表現になる動詞について、その具体的な検索意図表現との対応関係を記述している。さらに、「販売する」、「開発する」等の行動の時間的関係を示す動詞に対して、事象間の時間的な因果関係を記述する。この動詞シソーラスは、シソーラス辞書の１つの情報として用意する。

質問文中の動詞に対して、検索意図表現として対応する動詞をシソーラスから抽出して、動詞を展開する。また、動詞が時間的な因果関係を持つ場合はその関係によって、時間的に逆上る動詞にも展開する。

これらの動詞の操作によって、図１２のような質問例に応えることが可能になる。

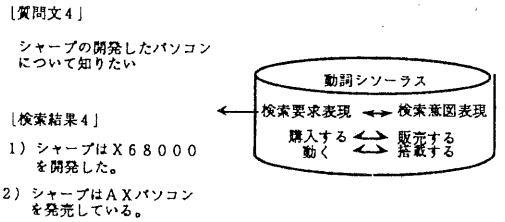


図１２　動詞シソーラスを利用した検索例

（2）範囲指定による検索

数値に対して「以上」や「以下」等の範囲指定に応じた検索を行う機能である。以下に本機能の質問例とその検索結果例を示す。

［質問文5］

100MIPS以上のマイクロ・プロセサは？

［検索結果5］

1）英Inmos Ltd.は、Transputer「T9000」のアーキテクチャを発表した。<文1>
ピーク性能は200MIPS。　<文2>

2）米Intelは、32ビット・マイクロプロセサi586の内部構造を明らかにした。<文1>
i586の性能は100MIPS以上。<文2>

●システム機能

質問文及びテキストの解析時に、名詞の属性の値を示す数詞文字列を、数値データに変換する数値処理を行なう。「以上」「以下」の範囲指定に応じて数値を比較し、該当する数値を含むテキストを検索する。

（3）事実ベース検索

検索ユーザが持っている知識を利用して検索内容を広げて検索を行う機能である。以下に本機能の質問例とその検索結果例を示す。

［質問文6］

廉価版のi486について知りたい。

［検索結果6］

1）米Intel Corp.は廉価版のi486を開発中である。

2）米i486SXを搭載した初めてのマシンである。

●システム機能

検索ユーザが検索の際に、まず「廉価版のi486はi486SX」というユーザが持っている知識を入力する。システムはその入力文を質問文や対象テキストと同様に構造化し、検索ユーザ用の事実ベースの中に格納する。次に、質問文の意味構造を抽出する。その際に、事実ベースを参照する。そこで、「廉価版のi486はi486SX」という情報を参照し、＜廉価版のi486＞という構造を＜i486SX＞にも展開する。この展開した構造で対象テキストとの照合を行なう。この展開構造を用いて、［検索結果6］の2）の文を含む記事を検索する。

（4）連続検索支援機能

直前の検索結果を基に、さらに検索要求がおきる場合の連続した検索を効率良く行うための支援機能である。直前の検索結果の中から検索者が1文あるいは文の一部を選び、それを利用して検索する。

以下に、本機能の質問例とその検索結果例を示す。

［質問文7］

ｉ４８６について知りたい。

［検索結果7］

1）米intelが廉価版のi486を開発している。

2）米intelは、i486の上位に位置する32ビットマイクロプロセサを発表した。<文1>
→［質問文8］
型名はi586になる。　<文2>

［検索結果8］

米intelは、32ビットマイクロプロセサであるi586の内部構造を明らかにした。

●システム機能

直前の検索結果の中で検索ユーザの指定した文から検索要求構造を抽出して、連続した検索を効率良く行う。

以下に上記検索例の処理の流れを、図13に沿って示す。

・［質問文7］を解析して、意味構造を抽出し、対象テキストの意味構造と比較・照合して、［検索結果7］を得る。
・次に、［検索結果7］の＜文2＞を次の検索の質問文として指定する。
・上記の［検索結果7］の＜文2＞の「米intelは、i486の上位に位置する３２ビットマイクロプロセサを発表した。」という文を［質問文8］とする。その文の意味構造は

すでに存在しているので、その構造から検索要求構造を得る。

・上記の検索要求構造から［検索結果８］を検索することができる。

この例では［検索結果８］の得る為に、事実ベース検索機能も利用している。すなわち、検索ユーザが、第１回目の［検索結果７］の＜文１＞と＜文２＞より、「i586はi486の上位に位置する３２ビットマイクロプロセサである。」という情報を事実ベースに追加することを前提としている。

２回目の検索の際に、追加した事実ベースの情報を参照し、図１３の網掛けの構造部分を＜i586＞にも展開する。この展開した意味構造によって、［検索結果８］を検索することを可能にしている。

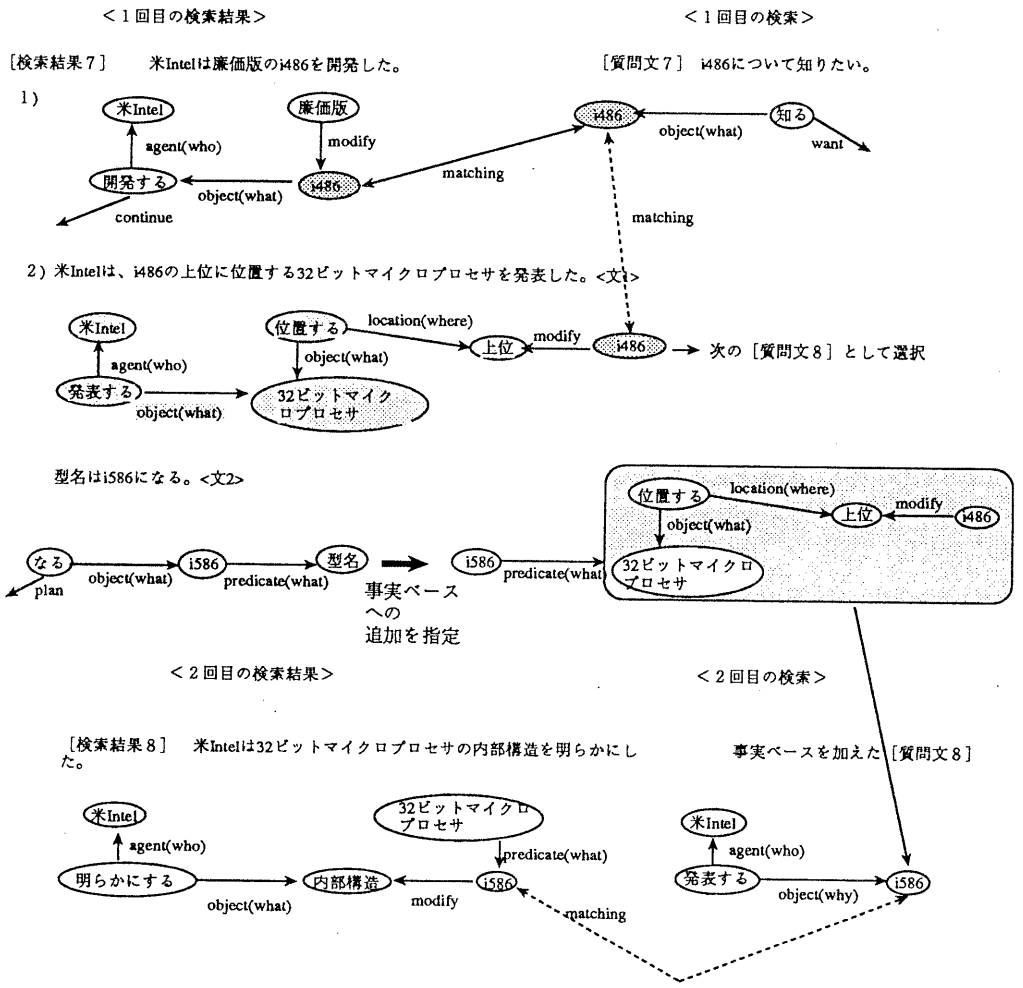


図１３　連続検索例

4.おわりに

フルテキストの日本語ニュースを対象に質問文の意図を反映し検索するシステムについて、システム機能を中心に述べた。現在、5記事(約70文)を対象に、小規模な機能実験システムを試作し、検索実験を行なっている。実験より、「関連情報検索機能」については、質問文の内容を理解し、質問文と同等な情報の検索に有効な機能であることを確認した。

現在の実験システムは小規模で、質問文の表現も限定されている。今後、さらにシステムの整備を計り、残りの機能についてもその有効性を評価していく。また、対象記事や質問の内容を増やした場合、知識を追加した場合の機能の有効性について評価していく予定である。

参考文献

1)五味淵亘：インデクシングの現場(3),JICSTにおけるインデクシングの実際、情報の科学と技術、Vol.39,No3,pp.99~109(1989).

2)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,pp.220~228(1991).

3)山口義一、杉山時之：国立国会図書館の雑誌記事索引システムにおける自然語による索引語自動抽出システムの概要とその索引語の分析、科学技術文献サース,Vol.32,No.4,pp.31~40(1988).

4)神尾達夫：新聞記事データベースにおけるキーワード自動抽出、情報管理、Vol.32,No.4, pp.283~293(1988)

5)根岸正光：フルテキスト・データベースの応用動向、情報処理、Vol.33,No.4,pp.413~420(1992)

6)石川徹也：知識データベースの構築と提供について－図書館の理念、図書館情報学（情報学）の立場から－、日本電子化辞書研究所ワークショップ論文集(TR-034),pp.94~100(1992).

7)Alberico,R & Micco,M.：Expert System for Reference and Information Retrieval, Meckler, p.395(1990).

8)Blair,D.C.：Language and Representation in Information Retrieval,Elsevier,p.335(1990).

9)米田健二：メタモルフ(METAMORPH)、テキストを理解する人工知能ソフトウエア、オンライン検索、Vol.10,No.3,pp.117~126(1989).

10)玉川義人、穂鷹良介：JDMFにおける役割と考察、情報処理学会データベースシステム研究会、P.10(1992).10