(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4282769号
(P4282769)
(45)発行日　平成21年6月24日(2009.6.24)　(24)登録日　平成21年3月27日(2009.3.27)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 170A
G06F 17/30 330C

請求項の数 3　(全 19 頁)

(21)出願番号	特願平10-538539
(86)(22)出願日	平成10年2月11日(1998.2.11)
(65)公表番号	特表2001-513243(P2001-513243A)
(43)公表日	平成13年8月28日(2001.8.28)
(86)国際出願番号	PCT/US1998/003005
(87)国際公開番号	WO1998/039714
(87)国際公開日	平成10年9月11日(1998.9.11)
審査請求日	平成17年2月9日(2005.2.9)
(31)優先権主張番号	08/886,814
(32)優先日	平成9年3月7日(1997.3.7)
(33)優先権主張国	米国(US)

(73)特許権者
マイクロソフト　コーポレイション
アメリカ合衆国　ワシントン州　９８０５２―６３９９　レッドモンド　ワン　マイクロソフト　ウェイ（番地なし）
(74)代理人
弁理士　谷　義一
(74)代理人
弁理士　阿部　和夫
(72)発明者　メセリー　ジョン　ジェイ
アメリカ合衆国　ワシントン州　９８１１０　ベインブリッジ　アイランド　オリンパス　ビーチ　ロード　９５１５

最終頁に続く

(54)【発明の名称】テキストの意味論的表現を利用した情報の検索

(57)【特許請求の範囲】
【請求項１】
第２のボディのテキストのあるパッセージに関連する第１のボディのテキストの複数のパッセージを特定するための、メモリ及び処理装置を備えるコンピュータ・システムによって実行される方法において、
前記第１のボディのテキストの前記複数のパッセージの各々に対して、前記処理装置が、パッセージを品詞で分解してパッセージ中の選択された複数の語の間の構文的な関係を特徴付ける第１の論理形式を構築するステップと、前記処理装置が、前記選択された複数の語の少なくともいくつかのための上位語含むように、構築された前記第１の論理形式を拡張するステップと、前記処理装置が、前記選択された複数の語の各々から、前記選択された複数の語の各々に対応する前記第１のボディのテクスト中の箇所へのマッピングを前記メモリのインデックスに格納ステップと、前記処理装置が、前記選択された複数の語の前記少なくともいくつかの上位語の各々から、前記上位語の各々に対応する前記第１のボディのテクスト中の箇所へのマッピングを前記メモリのインデックスに格納するステップと、
前記処理装置が、前記第２のボディのテキストの前記あるパッセージを品詞で分解して前記あるパッセージ中の選択された複数の語の間の構文的な関係を特徴付ける第２の論理形式を構築するステップと、
前記処理装置が、前記あるパッセージ中の前記選択された複数の語の少なくともいくつかのための上位語を含むように、構築された前記第２の論理形式を拡張するステップと、

前記処理装置が、前記インデックスに格納された語のうち、前記第２のボディのテキストの前記あるパッセージの前記選択された複数の語の各々に対応する語、または、前記あるパッセージの前記選択された複数の語の前記少なくともいくつかのための上位語の各々に対応する語を特定して、前記第２のボディのテキストの前記あるパッセージに関連する第１のボディのテキストのパッセージを特定するステップと
を備えることを特徴とする方法。
【請求項２】
前記第１のボディのテキストは複数の文書から成っており、
前記第１のボディのテキストの前記複数のパッセージの各々の箇所は、前記複数の文書のうちの各パッセージを含む文書の文書番号を含むことを特徴とする請求項１に記載の方法。
【請求項３】
第２のボディのテキストのあるパッセージに関連する第１のボディのテキストの複数のパッセージを特定するのに適合されたコンピュータ・システムであって、
前記第１のボディのテキストの前記複数のパッセージの各々に対して、パッセージを品詞で分解してパッセージ中の選択された複数の語の間の構文的な関係を特徴付ける第１の論理形式を構築し、前記選択された複数の語の少なくともいくつかのための上位語を含むように構築された前記第１の論理形式を拡張し、前記選択された複数の語の各々から、前記選択された複数の語の各々に対応する前記第１のボディのテクスト中の箇所へのマッピングをインデックスに格納し、前記選択された複数の語の前記少なくともいくつかの上位語の各々から、前記上位語の各々に対応する前記第１のボディのテクスト中の箇所へのマッピングを前記メモリのインデックスに格納するインデックス付けコンポーネントと、
前記第２のボディのテキストの前記あるパッセージを品詞で分解して前記あるパッセージ中の選択された複数の語の間の構文的な関係を特徴付ける第２の論理形式を構築し、前記あるパッセージ中の前記選択された複数の語の少なくともいくつかのための上位語を含むように、構築された前記第２の論理形式を拡張することによって、前記第２のボディのテキストの前記あるパッセージ中の選択された複数の語の間の意味的な関係を特徴付けるように適合された意味的関係特徴化コンポーネントと、
前記インデックスに格納された語のうち、前記第２のボディのテクストの前記あるパッセージの前記選択された複数の語の各々に対応する語、または、前記あるパッセージの前記選択された複数の語の前記少なくともいくつかのための上位語の各々に対応する語を特定して、前記第２のボディのテキストの前記あるパッセージに関連する第１のボディのテキストのパッセージを特定する関連パッセージ特定コンポーネントと
を備えることを特徴とするコンピュータ・システム。
【発明の詳細な説明】
技術分野
本発明は情報検索の分野に関し、特に情報検索のトークン化の分野に関する。
発明の背景
情報検索とは照会または照会文書中に目標文書語が出現するとそれを識別するプロセスのことである。情報検索は明示的なユーザー探索照会の処理、特定の文書に関連する文書の特定、２つの文書の類似性の判定、文書の特徴の抽出、および文書の要約を含む、幾つかの状況に適用して好結果を得ることができる。
情報検索には標準的には２段階のプロセスが含まれている。すなわち（１）索引作成段階では文書は先ず（ａ）文書中の各々の語を“トークン”と呼ばれる情報検索エンジンか理解でき、これによって区別されることができる一連の文字へと変換するプロセス（文書の“トークン化（tokenizing）”として知られている）、および（ｂ）各々のトークンからトークンが出現した文書中の箇所への索引図を作成するプロセスと、（２）照会段階では、照会（または照会文書）が同様にトークン化され、かつ索引と比較されて、トークン化された照会でトークンが出現する文書中の箇所を特定するプロセスである。
図１は情報検索プロセスを示した概略的なデータの流れ図である。索引作成段階では、目

標文書１１１がトークナイザ１１２へと提出される。目標文書は文章のような多数のストリングからなっており、その各々は目標文書の特定箇所に出現する。目標文書中のストリングとそれらの語の箇所がトークナイザ１２０に送られ、このトークナイザによって各ストリング中の語は情報検索エンジン１３０が理解でき、これによって識別可能な一連のトークンへと変換される。情報検索エンジン１３０の索引構成部１３１がトークンとその箇所を索引１４０に追加する。索引は各々の一意的なトークンをそれが目標文書中で出現した箇所へとマッピングする。必要ならば、多数の異なる目標文書を索引に加えるために、このプロセスを繰り返してもよい。このように索引１４０が多数の目標文書中のテキストを表す場合、箇所情報には好適には各々の箇所ごとに、この箇所に対応する文書の表示が含まれている。

照会段階では、テキスト照会１１２がトークナイザ１２０に提出される。この照会は単一のストリングでも文章でもよく、または文書全体が多数のストリングから構成されていてもよい。トークナイザ１２０は照会テキストト１１２中の語を、目標文書をトークンに変換したと同じ態様でトークンに変換する。トークナイザ１２０はこれらのトークンを情報検索エンジン１３０の索引検索部１３２へと送る。情報検索エンジンの索引検索部は索引１４０を探索して目標文書中にトークンが出現しているかどうかを検索する。情報検索エンジンの情報検索部は、各々のトークン毎に目標文書中のトークンが出現した箇所を特定する。このような箇所のリストが照会結果１１３として戻される。

従来形のトークナイザには標準的には、各々の大文字の小文字への変更、入力されたテキスト中の個々の語の識別、および語から接尾語を削除すること等のような入力されたテキストの外面的な変換が含まれている。例えば、従来形のトークナイザは下記のような入力されたテキスト・ストリング、父親が赤ちゃんを抱いている。（The father is holding the baby.）を下記のトークンへと変換するであろう。

the
father
is
hold
the
baby

このようなトークン化の方法はこれに基づいて、語の意味が照会テキスト中で意図した意味とは異なるような語の出現（occurrence）を過剰に含める探索を行う傾向がある。例えば、サンプルで入力されたテキスト・ストリングは“抱く（hold）”という動詞を、“支える、または抱きしめる”という意味で用いている。しかし、“hold”というトークンは“船舶の積荷領域”を意味する“hold”という語の用例と適合することがある。このトークン化の方法は更に、語が互いに照会テキスト中の語とは互いに異なる関係にある語の出現を過剰に含める傾向がある。

例えば、“父親”（father）が“抱く（held）”の主語であり、“赤ちゃん（baby）”が目的語である上記の入力されたテキスト・ストリングのサンプルは、“赤ちゃん”が主語で目的語ではない“父親と赤ちゃんが玩具を握っていた。”（The father and the baby held the toy.）と適合することがある。この方法は更に、照会テキストの代わりに、異なってはいるが意味的に関連する語を使用するテキスト・ストリングの出現を見落とすことがある。例えば、上記の入力されたテキスト・ストリングは“両親が赤ちゃんを抱いている”というテキスト・ストリングとは適合しないであろう。従来形のトークン化の上記のような欠点を考慮すると、トークン化されたテキスト中で明示されている意味的な関係をエンコードするトークナイザには重要なユーティリティがあるであろう。

発明の大要

本発明は入力されたテキストを分解して論理形式を特定し、次に上位語（hypernym）を用いて論理形式を拡張する改良形のトークナイザを使用して情報検索を実施することを指向している。本発明を従来形の情報検索の索引構造および照会と組合わせて利用すれば、異なる意味を意図し、かつ語が互いに異なる関係を有している、識別されるテキスト・スト

リングの出現数が減少し、かつ異なってはいるが意味的に関連する用語が用いられている、識別されるテキスト・ストリングの出現数が増加する。

本発明は索引付きテキストと照会テキストの双方を分解してこの入力されたテキストの字句的、統語的、および意味論的な解析を行うことによって従来形のトークン化に伴う問題点を克服するものである。この分解プロセスは照会テキスト中の主要な役割を果たす語とその意図された意味とを識別し、かつこれらの語の相互関係を識別する１つ、または複数の論理形式を作成する。パーザは好適には入力されたテキストの深い（実際の）主語、動詞および深い（実際の）目的語に関連する論理形式を作成する。例えば、入力されたテキストが“父親が赤ちゃんを抱いている。”である場合、パーザは下記のような論理形式を作成する。

深い主語	動詞	目的語
父親	抱く	赤ちゃん

パーザは更に、これらの語に前記入力されたテキストで用いられている特定の意味を付与する。

ある語の特定の意味について、語の意味にとって総称的な用語（“上位語”）である他の語の意味を特定するディジタル辞書または類語辞典（“言語知識ベース”としても知られている）を用いて、本発明はパーザによって作成された論理形式に含まれる語をその上位語に変更して、これらのオリジナルの論理形式の意味に対して上位概念である全体的な意味を有する追加の論理形式を作成する。例えば、“両親”（parent）の意味が“父親”に付与された意味の上位語であり、“触れる”（touch″）の意味が“抱く”に付与された意味の上位語であり、“子供”（child）の意味と“人”（person）の意味が“赤ちゃん”に付与された意味の上位語であるという辞書の表示に基づいて、本発明は下記のように追加の論理形式を作成する。

深い主語	動詞	目的語
両親	抱く	赤ちゃん
父親	触れる	赤ちゃん
両親	触れる	赤ちゃん
父親	抱く	子供
両親	抱く	子供
父親	触れる	子供
両親	触れる	子供
父親	抱く	人
両親	抱く	人
父親	触れる	人
両親	触れる	人

次に本発明は、生成された全ての論理形式を情報検索システムが理解できるトークンへと変換し、この情報検索システムがトークン化された照会を索引と比較し、これらのトークンを情報検索システムに提出する。

【図面の簡単な説明】

図１は情報検索プロセスを示す概略的なデータの流れ図である。

図２はそれによって機構が好適に機能する汎用コンピュータ・システムの高レベルのブロック図である。

図３は目標文書を意味論的に表す索引を構成し、これにアクセスするために機構が好適に実施するステップを示す概略的な流れ図である。
図４は入力された文章用のトークンを生成するために機構が利用するトークン化ルーチンを示す流れ図である。
図５は簡単な論理形式を示す論理形式図である。
図６は機構がそのために図５に示した論理形式を構成する、入力されたテキスト断片を示す入力テキスト図である。
図７Ａは言語知識ベースによって識別される総称的関係のサンプルを示す言語知識ベース図である。
図７Ｂは一次論理形式の深い主語の上位語の選択、男（意味２）を示す言語知識ベース図である。
図８は一次論理形式の深い動詞の上位語の選択、キス（意味１）を示す言語知識ベース図である。
図９および図１０は一次論理形式の深い目的語の上位語の選択、子豚（意味２）を示す言語知識ベース図である。
図１１は拡張された論理形式を示す論理形式図である。
図１２は拡張された一次論理形式を置換することによって作成された派生的論理形式を示す図表である。
図１３は索引のサンプル内容を示す索引図である。
図１４は“男が馬にキスしている。”という照会に関して機構が好適に構成する論理形式を示す論理形式図である。
図１５は上位語を用いた一次論理形式の拡張を示している。
図１６は照会論理形式の深い目的語の総称の選択、馬（意味１）を示す言語知識ベース図である。
図１７は深い主語と動詞だけを含む部分的照会に対応する部分的論理形式を示す部分論理形式図である。
図１８は動詞と深い目的語だけを含む部分的照会に対応する部分的論理形式を示す部分論理形式図である。
発明の詳細な説明
本発明はテキストの意味論的な表現を用いた情報検索を行うことを指向している。従来形の情報検索のための索引構造および照会と組合わせて利用すれば、本発明によって、異なる意味を意図し、かつ語が互いに異なる関係を有している、識別されるテキスト・ストリングの出現数が減少し、かつ異なってはいるが意味論的に関連する用語が用いられている、識別されるテキスト・ストリングの出現数が増加する。
好適な実施例では、図１に示した従来形のトークナイザの代わりに改良形の情報検索トークン化機構（“機構”）が使用され、これは入力されたテキストを分解して論理形式を識別し、次に上位語を用いて論理形式を拡張する。本発明は索引付きテキストと照会テキストの双方を分解してこの入力されたテキストの字句的、統語的、および意味論的な解析を行うことによって従来形のトークン化に伴う問題点を克服するものである。この分解プロセスは照会テキスト中の主要な役割を果たす語とその意図された意味とを識別し、かつこれらの語の相互関係を識別する１つ、または複数の論理形式を作成する。パーザは好適には入力されたテキストの深い主語、動詞および深い目的語に関連する論理形式を作成する。例えば、入力されたテキストが“父親が赤ちゃんを抱いている。”である場合、バーザは深い主語が“父親”であり、動詞が“抱く”であり、深い目的語が“赤ちゃん”であることを示す論理形式を作成する。入力されたテキストを論理形式へと変換することによって、修飾語が削除され、かつ時制と態（能動、受動）の差が無視されることで入力されたテキストはその根底的な意味へと抽出されるので、入力されたテキスト・セグメントを論理形式に変換することによって、同じ概念を表現するために自然語で用いられることがある多くの異なる表現方法が統一される傾向がある。
ある語の特定の意味について、語の意味にとって総称的な用語（“上位語”）である他の

語の意味を特定するディジタル辞書または類語辞典（“言語知識ベース”としても知られている）を用いて、本発明はパーザによって作成された論理形式に含まれる語をその上位語に変更して、これらのオリジナルの論理形式の意味に対して上位概念である全体的な意味を有する追加の論理形式を作成する。次に本発明は、生成された全ての論理形式を情報検索システムが理解できるトークンへと変換し、この情報検索システムがトークン化された照会を索引と比較し、これらのトークンを情報検索システムに提出する。

図２はそれによって機構が好適に機能する汎用コンピュータ・システムの高レベルのブロック図である。コンピュータ・システム２００は中央処理装置（ＣＰＵ）２１０と、入力／出力装置２２０と、コンピュータ記憶装置（メモリ）２３０とを含んでいる。入力／出力装置の間にはハードディスク駆動装置のような記憶装置２１１が配置されている。入力／出力装置は更にコンピュータによる読出し可能な媒体駆動装置２２２を含んでおり、これを用いてＣＤ－ＲＯＭのようなコンピュータによる読出し可能な媒体に備えられている機構を含むソフトウェア製品を据えつけることができる。入力／出力装置は更に、コンピュータ・システム２００がインターネットを介して他のコンピュータ・システムと通信できるようにするインターネット接続２２３をも含んでいる。好適には機構２４０からなるコンピュータ・プログラムはメモリ２３０内に備えられ、ＣＰＵ２１０上で実行する。機構２４０は更に、論理形式の語に意味番号を付与するためにパーザによって利用される言語知識ベース２４２を含んでいる。この機構は更に、言語知識ベースを利用して生成された論理形式の語の上位語を識別する。好適にはメモリ２３０は更に目標文書から生成されたトークンから目標文書中の箇所へとマッピングするための索引２５０を含んでいる。メモリ２３０は更に索引２５０中の目標文書から生成されたトークンを記憶し、かつ照会から生成されたトークンと適合するトークンを索引から識別するための情報検索エンジン（“ＩＲエンジン”）２６０をも含んでいる。前記機構は上記のように構成されたコンピュータ・システムで好適に実施されるが、異なる構造のコンピュータ・システムでも実施できることが当業者には理解されよう。

図３は目標文書を意味的に表す索引を構成し、かつこれにアクセスするために前記機構によって好適に実行されるステップの概略的な流れ図である。簡略に述べると、この機構は先ず目標文書の各々の文章（単数または複数）を、同じ意味の上位語を含む文章中の重要語の相互の関係を示す拡張された論理形式を表す多数のトークンへと変換することによって、目標文書の意味的な索引を作成する。

機構は文章が出現する目標文書中の箇所と共にこれらの“意味的トークン”を索引中に記憶する。全ての目標文書の索引作成が完了した後、機構は索引に対する情報検索の照会を処理することができる。受理されたこのような照会の各々について、機構は、それが目標文書から文章をトークン化したと同様に、すなわち照会テキスト用の拡張された論理形式を共に表す意味的なトークンへと文章を変換することによって、照会のテキストをトークン化する。次に機構はこれらの意味的なトークンを索引中に記憶された意味論的なトークンと比較して、これらの意味論的なトークンが記憶れている目標文書の箇所を特定し、かつ照会との関連性を特定するためにこれらの意味論的なトークンを含む目標文書をランク付けする。この機構は好適には新たな目標文書のための意味論的なトークンを随時含めるように索引を更新する。

図３を参照すると、ステップ３０１－３０４で、この機構は目標文書中の各文章を巡回する。ステップ３０２で、本機構は図４に示したように文章をトークン化するルーチンを呼び出す。

図４は入力された文章またはその他の入力されたテキスト・セグメントのためのトークンを生成するために本機構が利用するトークン化ルーチンを示した流れ図である。ステップ４０１で、本機構は入力されたテキスト・セグメントから一次論理形式を構成する。前述したように、論理形式は文章または文章の断片の根底的な意味を表すものである。論理形式は入力されたテキスト・セグメントがパーザ２４１（図２）によって語句的、および意味論的に分解されることによって作成される。入力されたテキスト・ストリングを表す論理形式の構造に関する詳細な説明は、本明細書に参考文献として引用されている米国特許

出願第０８／６７４，６１０号を参照されたい。
本機構によって利用される論理形式は好適には文章の主要な動詞と、動詞の実際の主語（“深い主語”）である名詞と、動詞の実際の目的語（“深い目的語”）である名詞とを分離する。図５は一次論理形式のサンプルを示す論理形式図である。論理形式は３つの要素を有している。すなわち、深い主語の要素５１０と、動詞の要素５２０と、深い目的語の要素５３０である。論理形式の深い主語は“男”という語の意味２であることが判る。１つ以上の意味を有する語の場合の意味番号は、パーザが利用する言語知識ベースによって定義されたとおりに語に付与された特定の意味を示すものである。例えば、“man”という語は人を表す第１の意味と、大人の弾性を意味する第２の意味を有することができよう。
論理形式の動詞は“kiss”という語の第１の意味である。最後に、深い目的語は“pig”という語の第２の意味である。この論理形式の短縮バージョンはその第１の要素として深い主語を有し、第２の要素として動詞を有し、かつ第３の要素として深い目的語を有する順序付けされた３つの語である。
（man,kiss,pig）
図５に示した論理形式は多数の異なる文章と文章の断片を特徴付けるものである。例えば、図６は本機構がそのために図５に示した論理形式を構成する入力されたテキストを示す入力テキスト図である。図６は入力された文章の断片“mankissing a pig（子豚にキスしている男）”を示している。この語句は文書５の語番号１５０で出現し、語の位置１５０、１５１、１５２、および１５３を占めていることが判る。本機構がこの入力テキストの断片をトークン化すると、図５に示した論理形式が生成される。本機構は更に、入力された下記のテキスト・セグメントについても図５に示した論理形式を生成する。
The pig was kissed by an unusual man.（子豚が異常な男にキスされた。）
The man will kiss the largest pig.（男が一番大きい子豚にキスするつもりだ。）
Many pigs have been kissed by that man.（これまで多くの子豚がその男によってキスされた。）
前述したように、入力されたテキストを論理形式へと変換することによって、修飾語が削除され、かつ時制と態の差が無視されることで入力されたテキストはその根底的な意味へと抽出されるので、入力されたテキスト・セグメントを論理形式に変換することによって、同じ概念を表現するために自然語で用いられることがある多くの異なる表現方法が統一される傾向がある。
図４に戻ると、本機構が入力されたテキストから図５に示した論理形式のような一次論理形式を構成した後、機構はステップ４０２を継続して上位語を用いてこの一次論理形式を拡張する。ステップ４０２の後、トークン化ルーチンに戻る。
前述したように、上位語は特定の語と“ｉｓａ”（・・である）の関係を持っている属の語である。例えば、“車両”という語は“自動車”の上位語である。本機構は好適には言語知識ベースを用いて一次論理形式の語の上位語を識別する。このような言語知識ベースは標準的には語の上位語を識別する意味論的なリンクを含んでいる。
図７Ａは言語知識ベースによって識別される上位語の関係のサンプルを示す言語知識ベース図である。図７Ａは、それ以降の言語知識ベースと同様に、説明を容易にするために簡略化されており、本明細書の説明には直接関連がなる言語知識ベースに一般に見られる情報を省いていることに留意されたい。図７Ａの各々の上向きの矢印は語をその上位語に結び付けている。例えば、男（意味２）という語を人（意味１）という語７１４に結び付けた矢印があり、これは人（意味１）が男（意味２）の上位語であることを示している。逆に、男（意味２）は人（意味１）の“下位語”であるということができる。
一次論理形式をそれによって拡張する上位語を識別する際に、本機構は上位語の下位語との“同類性”（coherency）に基づいて一次論理形式の各々の語ごとに１つ、またはそれ以上の上位語を選択する。このようにして上位語を選択することによって、本機構は入力されたテキスト・セグメントの意味を越えて、しかし制御された分量だけ論理形式の意味を一般化する。一次論理形式の特定の語ごとに、本機構は一次論理形式の語に近い上位語

を選択する。例えば、図７Ａを参照すると、一次論理形式で出現する男（意味２）７から始まって、本機構はその上位語である人（意味１）７１４を選択する。次に本機構は、人（意味１）が最初の語である男（意味２）７１１に対して設定された同類の上位語を有しているか否かに基づいて、人（意味１）７１４、動物（意味３）７１５をも選択するか否かを判定する。最初の語（意味２）７１１以外の、人と言う語のあらゆる意味の多数の下位語が最初の語である男（意味２）７１１と少なくともしきい値レベルの同類性を備えている場合は、人（意味１）７１４は男（意味２）７１１に対して設定された同類の下位語を有している。

上位語の異なる意味の下位語どうしの同類性のレベルを判定するため、本機構は好適には言語知識ベースに諮ってこれらの語の文章どうしの同類性の度合いを示す同類性の重みを得る。図７Ｂは男（意味２）と、人（意味１）および人（意味５）の他の下位語との同類性の重みを示す言語知識ベースの図である。この図は、男（意味２）と女（意味１）との同類性の重みが“００７５”てあり、男（意味２）と子供（意味１）との重みが“００２９”であり、男（意味２）と悪役（villain）（意味１）との重みが“０００３”であり、男（意味２）と主役（意味７）（lead）との重みが“０００２”であることを示している。これらの同類性の重みは好適には一対の語の意味間の言語知識ベースによって保持される意味論的な関係のネットワークに基づいて、論理知識ベースによって計算される。言語知識ベースを利用した一対の語の意味間の同類性の重みの詳細な説明については、本発明に参考文献として引用されている米国特許出願第　　号（特許代理人の件番第６６１００５．５２４号）、「語間の同類性の判定」を参照されたい。

これらの同類性の重みに基づいて下位語の集合が類似しているか否かを判定するために、本機構はしきい値の数の同類性の重みが、同類性の重みのしきい値を超えているか否かを判定する。好適なしきい値百分率は９０％であるが、機構の性能を最適化するため、好適にはしきい値百分率を調整してもよい。更に機構の性能が最適化されるように同類性の重みのしきい値を構成してもよい。同類性の重みのしきい値は好適には言語知識ベースによって得られる同類性の重みの全体的な配分と調整される。ここでは“００１５”のしきい値が用いられることが示されている。従って本機構は最初の語と、上位語の全ての意味のその他の下位語との間の同類性の重みの少なくとも９０％が“００１５”である同類性の重みのしきい値、またはそれ以上にあるか否かを判定する。図７Ｂから、この条件は男（意味１）に対する人の下位語によっては満たされず、一方、男（意味１）と女（意味１）との間の、および男（意味１）と子供（意味１）との間の同類性の重みは“００１５”以上であり、男（意味１）と悪役（意味１）との間の、および男（意味１）と主役（意味７）との間の同類性の重みは“００１５”未満であることが判る。従って本機構はそれ以上の上位語・動物（意味３）７１５、または動物（意味３）のどの上位語をも選択しない。その結果、一次論理形式を拡張するために上位語・人（意味１）７１４が選択される。

一次論理形式を拡張するため、本機構は更に一次論理形式の動詞と深い目的語の上位語をも選択する。図８は一次論理形式の動詞・キス（意味１）の上位語の選択を示している。この図から、触れる（意味２）がキス（意味１）の上位語であることが判る。図は更にキス（意味１）と、触れるの全ての意味のその他の下位語との間の同類性の重みをも示している。本機構は先ず一次論理形式の動詞・キス（意味１）に近い上位語である触れる（意味２）を選択する。触れる（意味２）、相互に関係する（意味９）（interact）を選択するか否かを判定するため、本機構はキス（意味１）と触れるの全ての意味のその他の下位語との間のどの位の数の同類性の重みが、しきい値の同類性の重みと少なくとも同じ値であるかを判定する。これらの４つの同類性の重みのうち２つしか“００１５”である同類性の重みのしきい値と少なくととも同じ値ではないので、本機構は触れる（意味２）、互いに関係する（意味９）の上位語を選択することはない。

図９および図１０は一次論理形式の深い目的語の上位語および子豚（意味２１）の選択を示す言語知識ベースの図である。図９から、豚（swine）の唯一の意味の上位語の９０％以上（実際には１００％）が“００１５”の同類性の重みのしきい値にあるか、それに近いので、本機構は一次論理形式を拡張するために子豚（pig）（意味２）の上位語である

豚（意味１）（swine）、並びに豚（意味１）（swine）の上位語である動物（意味３）（animal）を選択することが判る。図１０から、動物の意味の下位語の９０％未満しか（実際には２５％）“００１５”、または約“００１５”の同類性の重みのしきい値にないので、本機構は継続して動物（意味）の上位語である生物を選択することはないことが判る。

図１１は拡張れた論理形式を示した論理形式図である。図１１から、拡張された論理形式の深い主語要素１１１０は男（意味２）という語１１１１に加えて上位語の人（意味１）１１１２を含んでいることが判る。動詞要素１１２０はキス（意味１）１１２１という語と共に上位語である触れる（意味２）１１２２を含んでいることが判る。更に、拡張された論理形式の深い目的語要素１１３０が、子豚（意味２）（pig）１１３１に加えて、上位語である豚（意味１）（swine）と動物（意味３）１１３２を含んでいることも判る。

拡張された論理形式の各要素において上位語をオリジナルの語で置換することによって、本機構は意味が一次論理形式に適正に近似する適正に多数の派生的論理形式を生成することができる。図１２は拡張された一次論理形式を置換することによって生成される派生的論理形式を示した図表である。図１２から、このような置換によって入力されたテキストの意味を各々が適正な正確さで特徴付ける１１の派生的論理形式が生成されることが判る。例えば、図１２に示された派生的な論理形式

（person,touch,pig）（人、触れる、子豚）

は、意味的に文章の断片、

man kissing a pig（男が子豚にキスしている）

に極めて類似している。

図１１に示した拡張された論理形式は一次論理形式プラス、これらの１１の派生的論理形式を表しており、これらは拡張された論理形式１２００としてよりコンパクトに表現されている。

（（男または人）、（キスまたは触れる）、（子豚または豚または動物））

本機構はこの拡張された論理形式から、従来形の情報検索エンジンによる処理が可能であるようにする論理トークンを生成する。最初に、本機構はある語が入力されたテキスト・セグメント中に深い主語、動詞または深い目的語のどれとして出現したかを識別する指定符号を、拡張された論理形式中の各語に添付する。それによって、“man（男）”という語が深い主語として照会用に入力されたテキスト・セグメント用の拡張論理形式に出現した場合、それが動詞であった拡張論理形式の一部として索引に記憶されている“man（人員を配置する）”という語と適合することが確実になくなる。論理形式の要素への指定符号のマッピングのサンプルは下記のとおりである。

論理形式要素	識別符号
深い主語	＿
動詞	
深い目的語	＃

指定符号のこのようなサンプル・マッピングを利用して、論理形式用に生成されたトークン“（男、キス、子豚）”には“男＿”、“キス＾”、および“子豚＃”が含まれよう。

従来形の情報検索エンジンによって作成された索引（複数）は一般に各々のトークンを目標文書のトークンが出現する特定箇所へとマッピングする。従来形の情報検索エンジンは例えば、文書番号を用いてこのような目標文書を表し、トークンの出現を含む目標文書を識別し、その目標文書中のトークンの出現箇所を特定する。このような目標文書の箇所を発見することによって、従来形の情報検索エンジンは、“ＰＨＲＡＳＥ”（語句用）演算子を用いた照会に応答して目標文書中に共に出現する語を識別することが可能であり、その際にＰＨＲＡＳＥ演算子が結び付ける語は目標文書中の近くにある必要がある。例えば、“赤い ＰＨＲＡＳＥ自転車”という照会は、文書５の語６１１の“赤い”という語、および文書５の語６１２の“自転車という語の出現と適合するが、文書７、部７６２の“赤い”という語、および文書７の語２０２の、“自転車”という語の出現とは適合しないであろう。目標文書の箇所を索引に記憶しておくことによって更に、従来の情報検索エン

ジンが照会に応答して、照会がなされたトークンが目標文書中に出現するポイントを特定することが可能になる。
目標文書の入力されたテキスト・セグメントから拡張された論理形式の場合は、拡張された論理形式のトークンが目標文書のその箇所で出現しない場合でも、本機構は好適には人工的な目標文書の箇所を各々のトークンに同様に割当てる。これらの目標文書の箇所を割当てることによって、（Ａ）従来形の探索エンジンがＰＨＲＡＳＥ演算子を利用して単一の一次、または派生的論理形式に対応する意味論的トークンの組合わせを識別することと、（Ｂ）本機構が割当てられた箇所を目標文書中の入力されたテキストの断片の実際の箇所と関連付けることの双方が可能になる。従って本機構は意味論的なトークンに以下のように箇所を割当てる。

論理形式要素	箇所
深い主語	（入力されたテキスト・セグメントの最初の語の箇所）
動詞	（入力されたテキスト・セグメントの最初の語の箇所）＋１
深い目的語	（入力されたテキスト・セグメントの最初の語の箇所）＋２

従って本機構は文書５、語１５０で始まる文章から導出された“男、キス、子豚）”について、拡張された論理形式のトークン用に目標文書の箇所を下記のように割当てる。すなわち、“男＿”および“人＿”、文書５、語１５０；“キス＾”、および“触れる＾”、文書５、語１５１、および“子豚＃”および“豚＃”および“動物＃”、文書５、語１５２である。
図３に戻ると、ステップ３０３で、本機構はトークン化ルーチンによって生成されたトークンをそれらが出現する箇所と共に索引に記憶する。図１３は索引のサンプル内容を示す索引図である。索引は各トークンから文書の特定、およびトークンが出現する文書中の箇所までをマッピングする。索引中のマッピングをより明解に示すために索引は表として示されているが、実際には索引は好適には木状形式のような、索引中のトークンの箇所をより効率よくサポートする多くの他の形式の１つで記憶されることに留意されたい。更に、索引のサイズを最小限にするため、接頭圧縮のような技術を用いて索引の内容を圧縮することが好適である。
ステップ３０３に基づいて、本機構は各々の語のためのマッピングを拡張論理形式で索引１３００中に記憶していることが判る。マッピングは深い主語である“男”および“人”から目標文書の文書番号５、語番号１５０までのマッピングが索引に記憶されている。語番号１５０は図６に示した入力されたテキスト・セグメントが開始される語の位置である。本機構は深い主語に対応するトークンに指定符号“＿”を添付したことが判る。この指定符号を添付することによって、本機構は後に索引を探索する際に、論理形式の動詞または深い目的語として出現するこれらの語の出現を検索せずに、論理形式の深い主語として出現するこれらの語の例だけを検索することができる。同様にして、索引は動詞の語である“キス”および“触れる”のトークンを含んでいる。これらの動詞語の入力によってこれらの語は目標文書の文書５、語番号１５１の箇所に、目標文書の深い主語の箇所の後に１語ずつマッピングされる。更に、指定符号“＾”がこれらの動詞語のためのトークンに添付されているので、これらの語がこのような形で出現しても、後に深い主語または深い目的語として出現したものとは見なされないことが判る。同様にくして、索引は深い目的語“動物”、“子豚”、および“豚”のためのトークンを含んでおり、これらの語は目標文書の文書番号５、語番号１５２の箇所に、すなわち語句（ＰＨＲＡＳＥ）が始まる目標文書の箇所から２語先の箇所にマッピングされる。深い目的語を索引中で深い目的語として識別するため、指定符号“＃”が深い目的語のためのトークンに添付される。索引が図示した状態にある場合、図１２に示した派生的な一次論理形式のいずれかについて索引を探索することによって、図６に示した入力されたテキストの断片を見出すことができる。
本機構が目標文書中に字句的に出現する語の、目標文書中のそれらの語の実際の箇所へのマッピングと、同じ索引中の目標文書の意味論的な表示の双方を記憶する好適な実施例では、索引中でアクセスされた場合に意味論的な表示の意味論的なトークンと字句的なトークンとを区別するために、意味論的表示の各々の意味論的トークンの語番号の値は、好適

にはいずれかの文書中の語の番号よりも大きい定数だけ増分される。図１３を簡略にするため、この定数の追加は図示していない。
この例では、本機構は拡張された論理形式の各々の語のためのトークンを索引に追加して、目標文書の意味論的な表示を形成する。しかし、好適な１実施例では、本機構はそれが索引に追加する拡張された論理形式のトークンの集合を、目標文書中の文書どうしを区別するのに有効であると思われる論理形式のトークンに限定する。索引に追加される拡張された論理形式のトークンの集合をこのように限定するため、本機構は好適には各トークンの逆文書頻度（Inverse Document Frequency）を判定する。その公式は下記の方程式（１）によって示されている。この実施例では、本機構はその逆文書頻度が最大しきい値を超えるトークンだけを索引に追加する。
図３に戻ると、トークンを索引に記憶した後、目標文書中の目下の文章の処理前にステップ３０４で、本機構は目標文書中の次の文章を処理するためにステップ３０１に戻って循環する。目標文書の全ての文章の処理が終了した後、本機構はステップ３０５に進行する。ステップ３０５で、本機構は照会のテキストを受理する。ステップ３０６－３０８で、本機構は受理した照会を処理する。ステップ３０６では、本機構はトークン化ルーチンを呼出して照会テキストをトークン化する。図１４はステップ４０１（図４）に従って“馬にキスする男”という照会のために本機構が好適に構成する論理形式を示す論理形式図である。この論理形式図から、深い主語が男（意味２）であり、動詞がキス（意味１）であり、深い目的語が馬（意味１）であることが判る。この一次論理形式は一次論理形式１４５０で、
（男、キス、馬）
としてより簡潔に表示される。
図１５はステップ４０２（図４）に基づいて上位語を用いた一次論理形式の拡張を示している。図１５から、目標文書からのサンプルの入力テキストと同様に、深い主語である男（意味２）が上位語である人（意味１）によって拡張され、動詞・キス（意味１）が上位語である触れる（意味２）で拡張されたことが判る。更に、深い目的語・馬（意味１）が上位語・動物（意味３）で拡張されたことが判る。
図１６は照会論理形式の深い目的語・馬（意味１）の上位語の選択を示す言語知識ベース図である。図１６から、動物（意味３）の上位語の９０％未満しか“００１５”である同類性の重みのしきい値、またはそれ以上にはないので、本機構は動物（意味３）の上位語である生物（意味１）を選択しないことが判る。
従って、本機構は論理形式を拡張するために上位語・動物（意味３）だけを用いるのである。
図３に戻ると、ステップ３０７で本機構は一次論理形式の語の意味の上位語を用いて構成された拡張論理形式１５５０（図１５）を用いて、適合するトークンが出現する目標文書中の箇所を索引箇所から検索する。本機構は好適には索引に下記の照会を発することによって上記の動作を行う。
（男＿または人＿）ＰＨＲＡＳＥ（キス＾または触れる＾）ＰＨＲＡＳＥ（馬＃または動物＃）
ＰＨＲＡＳＥ演算子は先行する演算数（オペランド）よりも１つだけ大きい語の箇所で後続の演算数の出現を突合わせる（match）。従って、照会によって深い主語である男＿または人＿が、深い目的語である馬＃または動物＃に先行する動詞キス＾または触れる＾に先行する箇所が突合わせされる。図１３の索引から、この照会が文書番号５、語番号１５０で満たされることが判る。
前記照会が索引中で満たされない場合は、本機構は異なる２つの部分的照会によって照会の提出を継続する。第１の部分的形式には深い主語と動詞だけが含まれ、目的語は含まれない。
（男＿または人＿）ＰＨＲＡＳＥ（キス＾または触れる＾）
図１７はこの第１の照会に対応する部分的論理形式を示した部分的論理形式図である。照会の第２の部分的形式には動詞と深い目的語が含まれるが、深い主語は含まれない。

（キス＾または触れる＾）ＰＨＲＡＳＥ（馬＃または動物＃）
図１８はこの第２の照会に対応する部分的論理形式を示した部分的論理形式図である。これらの部分的照会によって索引中の異なる深い主語または深い目的語を有する部分的論理形式の突合わせが行われ、また、深い主語または深い目的語を有していない部分的な論理形式が突合わせされよう。これらの部分的突合わせは、照会のための入力テキスト・セグメントと、代名詞の使用および暗示された深い主語および深い目的語を含む目標文書の入力テキスト・セグメントとの差を考慮に入れている。
図３に戻ると、索引中のトークンの適合の有無を識別した後、本機構は継続してステップ３０８で、一次論理形式または派生的論理形式に対応して、突合わせトークンの特定の組合わせが照会に対する関連性が高い順序で出現する目標文書のランク付けを行う。本発明の様々な実施例において、本機構は関連性に応じた文書のランク付けのための多数の公知のアプローチのうちの１つ、または複数の方法を利用し、それにはジャッカード（Jaccard）重み付けおよび２進項目インピーダンス重み付けが含まれる。本機構は好適には適合する目標文書をランク付けするために逆文書頻度と項目頻度待機の組み合わせを利用する。
逆文書頻度重み付けの特徴は、文書間でより少ない目標文書に出現するトークンの組合わせに、より大きい重みを付与する文書を区別するトークン組合わせの能力にある。例えば、写真の主題に関する目標文書群の場合、論理形式、
（写真家、フレーム、主題）
は、各文書群に出現する可能性があり、従って文書間を区別するための極めて良好な基準にはならないであろう。上記の論理形式は全ての目標文書に出現するので、その逆文書頻度は比較的少ない。トークンの組合わせの逆文書頻度の公式は下記のとおりである。

$$\text{逆文書頻度（トークンの組合わせ）}=\log\left(\frac{\text{目標文書の総数}}{\text{トークンの組合せを含む目標文書数}}\right)$$

（１）

文書中のトークンの組合わせの項目頻度の重み付けは、ある文書がトークンの組合わせ専用である度合いの尺度であり、かつ特定の照会トークンがより高頻度で出現する文書は照会トークンがより少ない頻度で出現する文書よりも関連性が高いものと想定されている。
文書中のトークンの組合わせの項目頻度の重みの公式は下記のとおりである。
項目頻度（トークンの組合わせ）＝文書中でトークンの組合わせが出現する数（２）
本機構は各々の突合わせ文書ごとにスコアを用いて文書をランク付けする。本機構は先ず下記の公式を用いて各文書中の各々の突合わせトークンの組合わせのスコアを計算する。
スコア（トークンの組合わせ、文書）＝逆文書頻度（トークンの組合わせ）ｘ項目頻度（トークンの組合わせ、文書）　（３）
次に本機構は下記の公式に基づいて各突合わせ文書中に適合するトークンの組合わせがあればその最高スコアを選択することによって、各突合わせ文書のスコアを計算する。

$$\text{スコア（文書）}=\text{最大}\left(\underset{\text{文書中のトークンの組合せ}}{\forall}\ \text{（スコア（トークンの組合せ、文書））}\right)$$

（４）

本機構が各文書についてスコアの計算を終了すると、本機構はこれらのスコーを増倍して、意味論的な突合わせとは別の照会の項目を反映するようにしてもよい。各文書ごとにスコアを増倍した後、必要ならば本機構は下記の公式に示すように文書のサイズを考慮に入れて各文書の正規化されたスコアを計算する。

$$\text{正規化されたスコア（文書）} = \frac{\text{スコア（文書）}}{\text{サイズ（文書）}} \qquad (5)$$

サイズ（文書）の項目は例えば文書中の文字数、語、または文書または文書の断片のような文書のサイズのいずれかの適正な尺度でよい。あるいは、余弦尺度による正規化、項目の重みの合計による正規化、および最高の項目重みによる正規化を含む、他の多くの正規化技術を用いて文書スコアを正規化してもよい。

各突合わせ文書ごとに正規化されたスコアを計算した後、本機構は文書の正規化されたスコア順に突合わせ文書のランク付けを行う。ユーザーは好適にはランク付けされた突合わせ文書の１つを選択して、その文書中の適合するトークンの箇所を探し、またはその文書の適合部分が表示されるようにしてもよい。

図３を参照すると、ステップ３０８で突合わせ目標文書のランク付けを行った後、本機構は好適にはステップ３０５で索引に対する次の照会テキストを受理する。

上記は突合わせトークンを含む文書の関連性によるランク付けを説明したものである。本発明の更に別の好適な実施例は同様に、適合が含まれる、関連性がある文書群と文書部分のそれぞれによってランク付けを行う。各々が１つ、またはそれ以上の文書を含む文書群へと編成された目標文書の場合は、本機構は好適には、更なる照会のために最も関連性が高い文書群を特定するために、適合が出現する文書群を関連性によってランク付けする。更に、本機構は好適には各々の目標文書を各部分に分割し、適合が出現する文書部分の関連性をランク付ける。これらの文書部分はある数のバイト、語、または文章を選択するか、または目標文書中に出現する構造的、書式的、または言語的なキューを用いて目標文書中で連続的に特定される。更に本機構は好適には特定のテーマに関する非連続的な文書部分を特定することもできる。

これまで本発明を特定の実施例を参考にして図示し、説明してきたが、本発明の範囲を離れることなく形式と細部の多くの変更または修正が可能てあることが当業者には理解されよう。例えば、トークナイザを直截に、論理形式構造の１つの語に各々が対応するトークンの代わりに、完全な論理形式構造に各々が対応するトークンを作成し、索引に記憶されるようにしてもよい。更に、意味論的な突合わせ成分を有する照会に他の種類の探索方法を組入れるために様々な公知の技術を適用してもよい。更に、照会には多数の意味論的な突合わせ成分が含まれるようにしてもよい。加えて、上位語以外に語間で識別される意味論的な関係性を利用して、一次論理形式を拡張してもよい。本機構は更に、前述のルーチンで字句知識ベースから上位語のリストを作成するのではなく、一次論理形式の各語について予め承認された代用可能な語のリストを利用して一次論理形式を拡張してもよい。更に、突合わせの精度を更に高めるため、トークナイザは語に特定された意味番号を語のためのトークン中でエンコードしてもよい。この場合は、上位語の集合の同類性のためのテストは、選択された上位語の全ての意味との類似性のテストよりも軽減される。１例では、人という語の意味１の上位語だけが男（意味２）という語の最初の意味との同類性のしきい値レベルにあればよい。可能性がある索引中の突合わせ項目には曖昧さが少ないので、誤った適合を生ずることがある項目の集合を制限することができる。このような理由から、論理形式の語と上位語の関係にある意味だけをテストすればよい。

【図１】

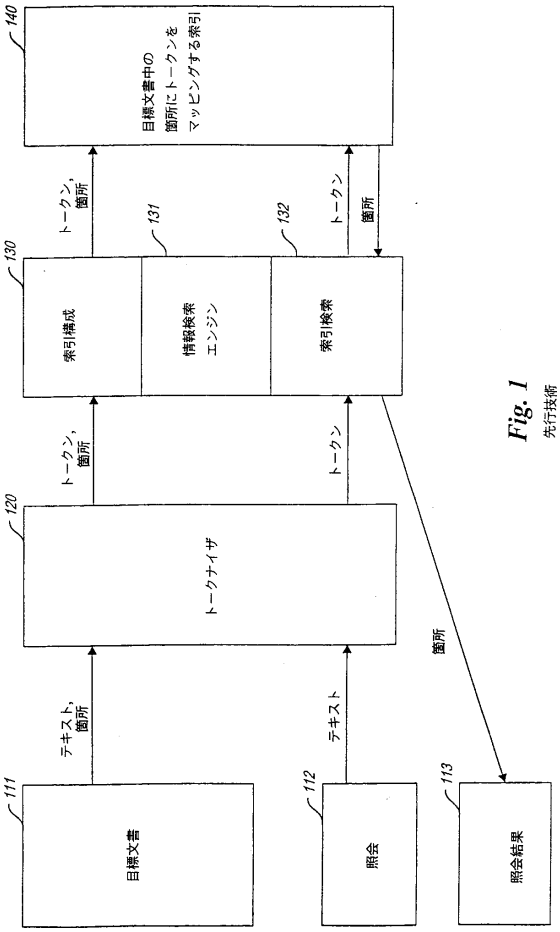

Fig. 1
先行技術

【図２】

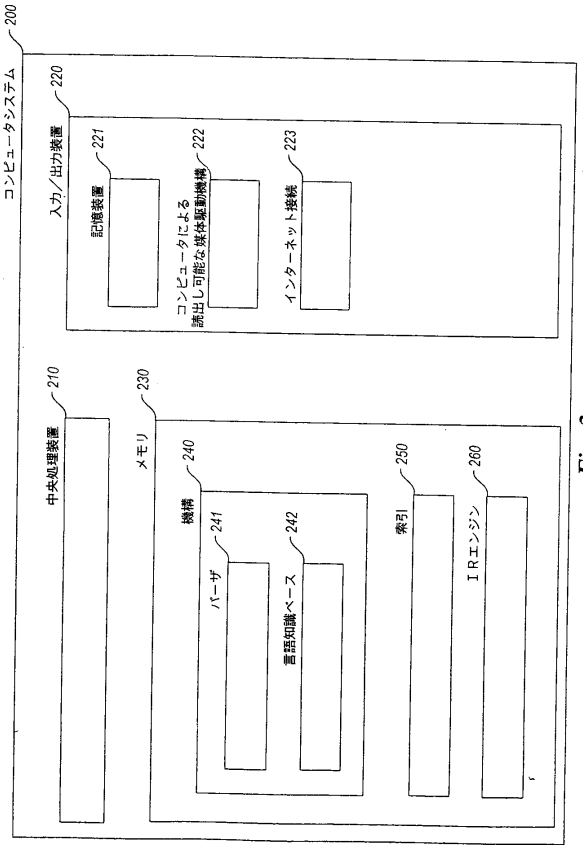

Fig. 2

【図３】

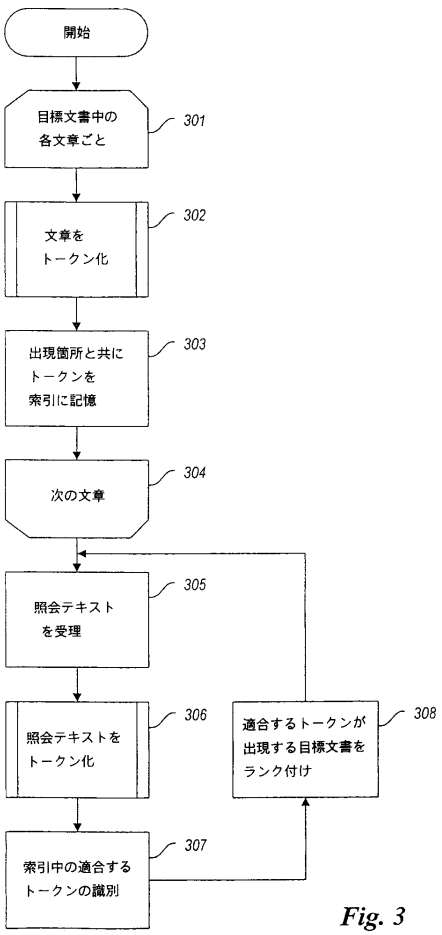

Fig. 3

【図４】

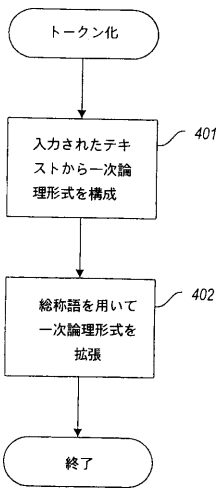

Fig. 4

【図５】

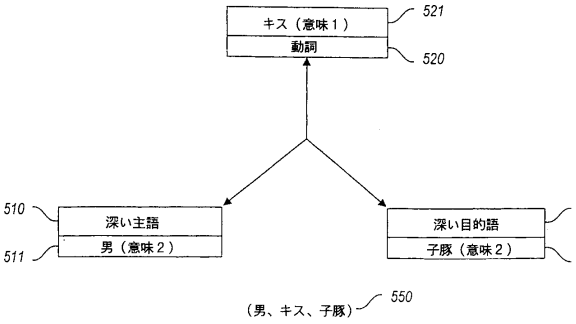

Fig. 5

【図６】

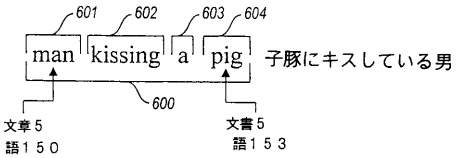

Fig. 6

【図７Ａ】

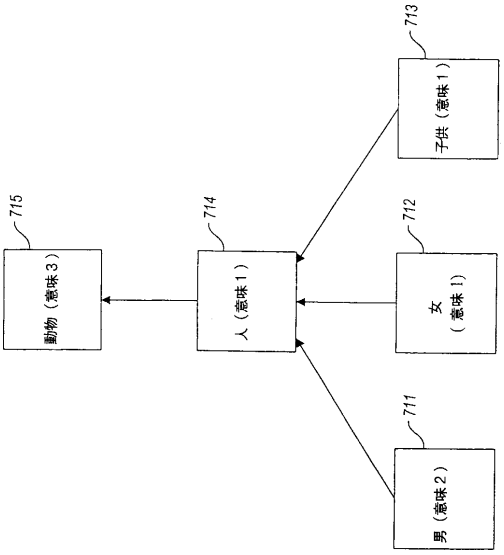

Fig. 7A

【図７Ｂ】

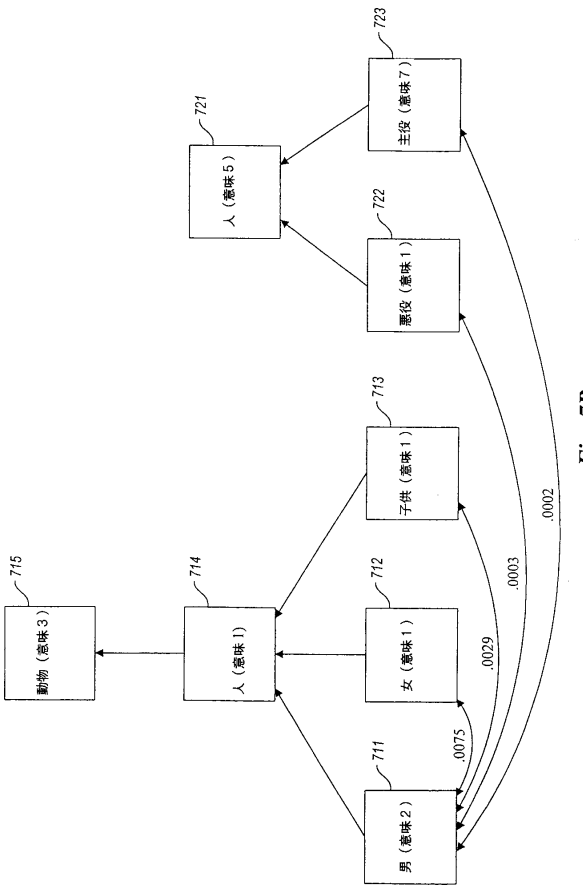

Fig. 7B

【図８】

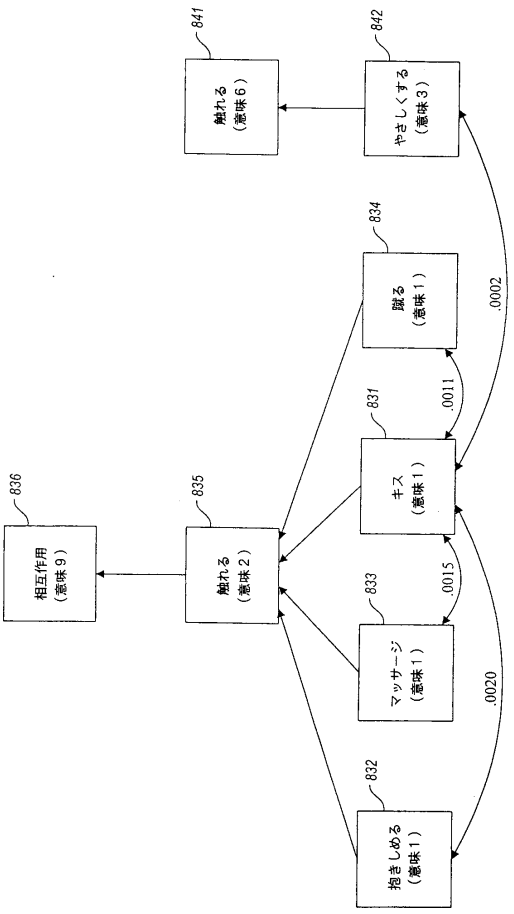

Fig. 8

【図９】

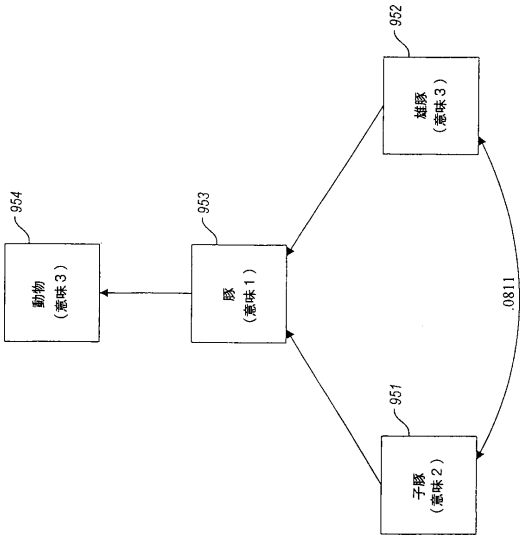

Fig. 9

【図１０】

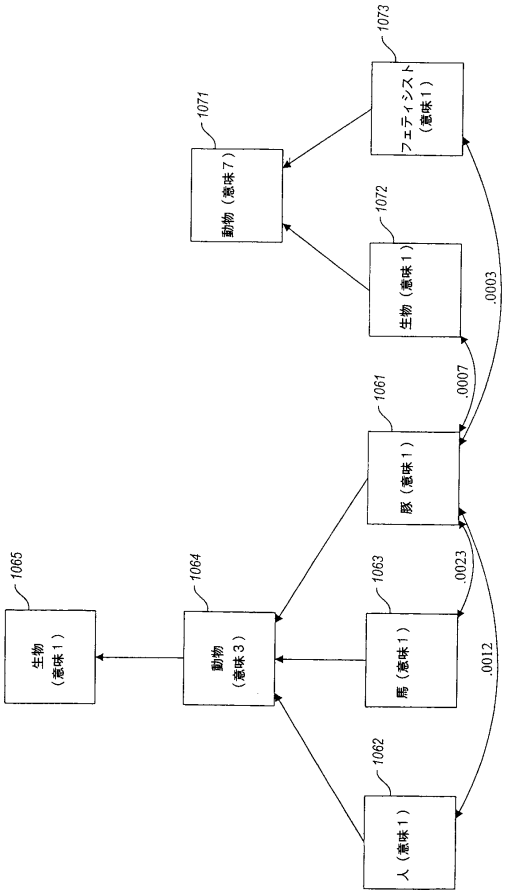

Fig. 10

【図１１】

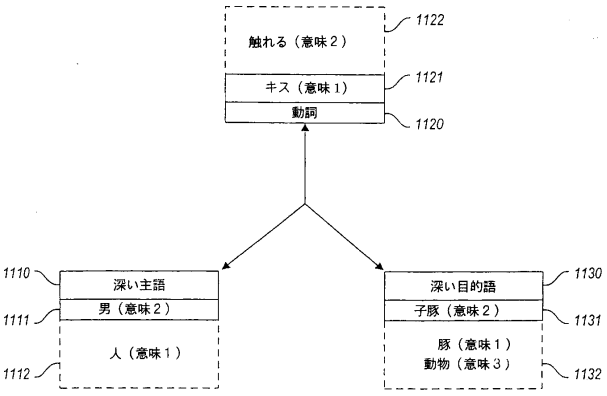

Fig. 11

【図１２】

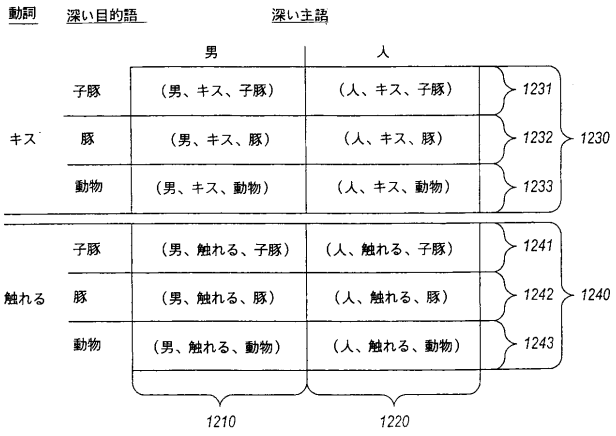

Fig. 12

【図１３】

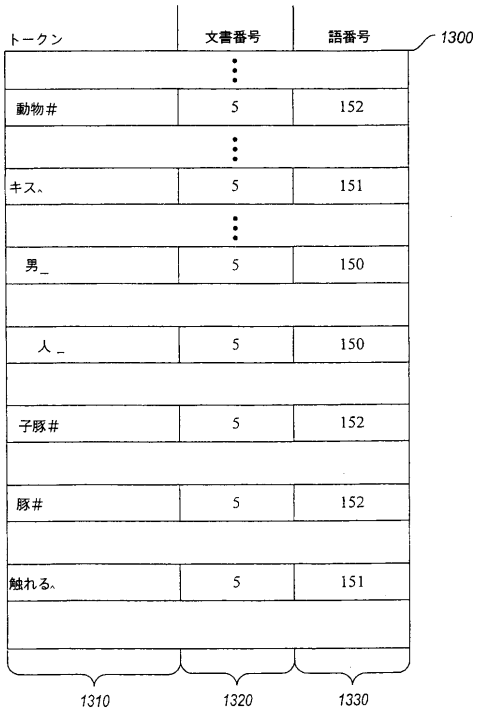

Fig. 13

【図１４】

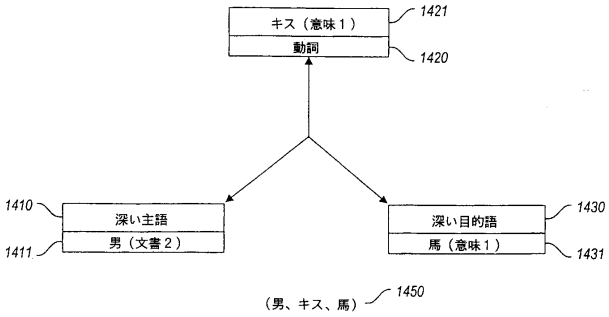

Fig. 14

【図１５】

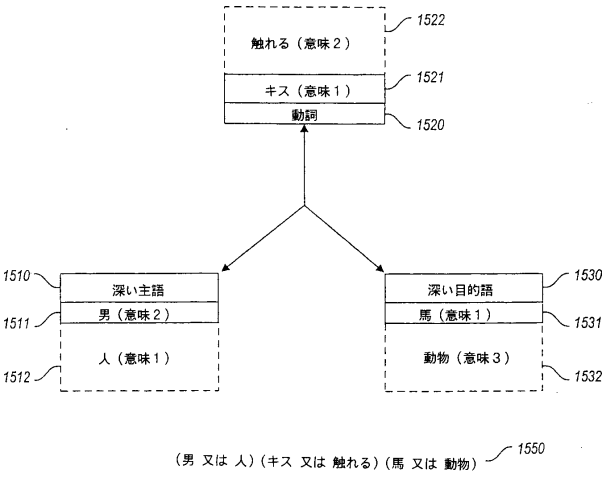

Fig. 15

【図１６】

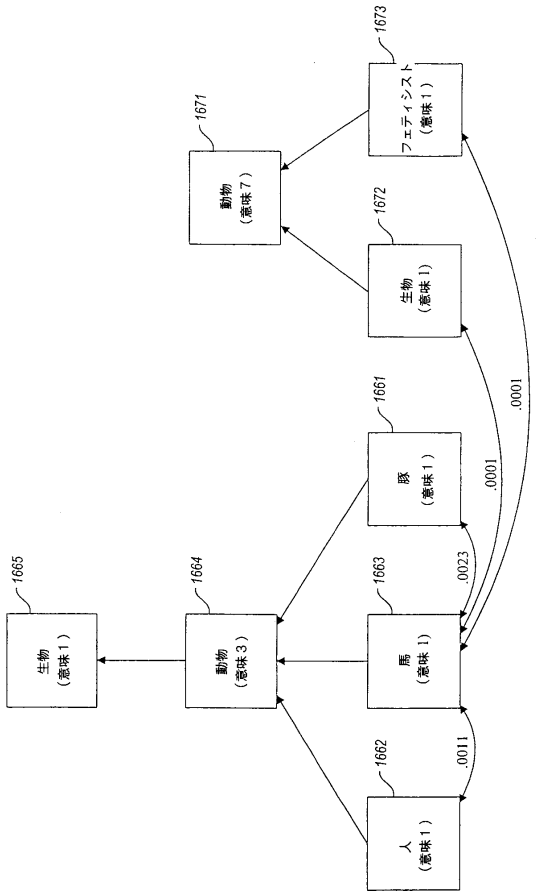

Fig. 16

【図１７】

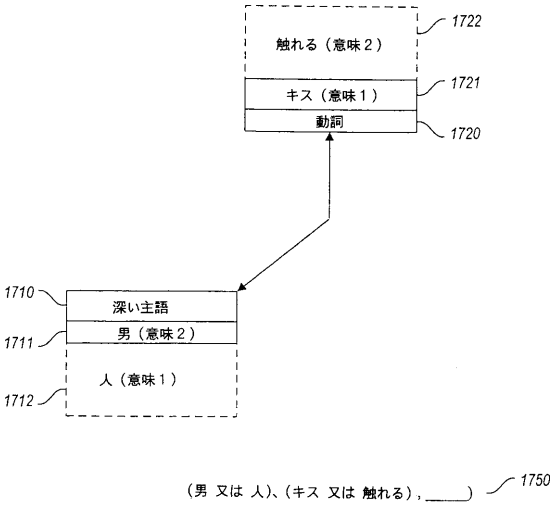

Fig. 17

【図１８】

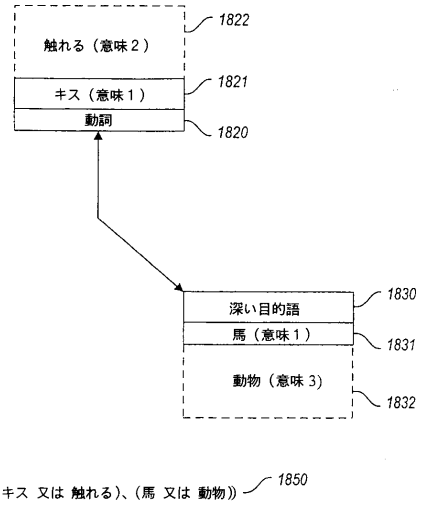

Fig. 18

フロントページの続き

(72)発明者　ヘイドーン　ジョージ　イー
アメリカ合衆国　ワシントン州　９８００８　ベルヴィュー　ワンハンドレッドアンドシックスティーフィフス　プレイス　ノースイースト　３２１１
(72)発明者　リチャードソン　スティーブン　ディー
アメリカ合衆国　ワシントン州　９８０５２　レッドモンド　ノースイースト　ワンハンドレッドアンドサーティセカンド　１８０２８
(72)発明者　ドーラン　ウィリアム　ビー
アメリカ合衆国　ワシントン州　９８０５２　レッドモンド　ノースイースト　ワンハンドレッドアンドフィフティーサード　コート　７４１２
(72)発明者　ジェンセン　カレン
アメリカ合衆国　ワシントン州　９８００８　ベルヴィュー　ワンハンドレッドアンドシックスティーフィフス　プレイス　ノースイースト　３２１１

審査官　鈴木　和樹

(56)参考文献　特開平０６－０１２４５１（ＪＰ，Ａ）
特開平０３－１７２９６６（ＪＰ，Ａ）
特開平０６－３０９３６２（ＪＰ，Ａ）
特開昭６４－０５５６４２（ＪＰ，Ａ）
特開平０３－０８７９７５（ＪＰ，Ａ）
特開平０５－２５０４１３（ＪＰ，Ａ）
特開平０７－０４４５６７（ＪＰ，Ａ）
特開平０８－００６９７１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/27 - 30
NRIサイバーパテント