(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-228033
(P2005-228033A)
(43)公開日　平成17年8月25日(2005.8.25)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	210A	5B009
G06F 17/21	G06F 17/30	170A	5B075
	G06F 17/30	320C	
	G06F 17/21	550A	
	G06F 17/21	590E	

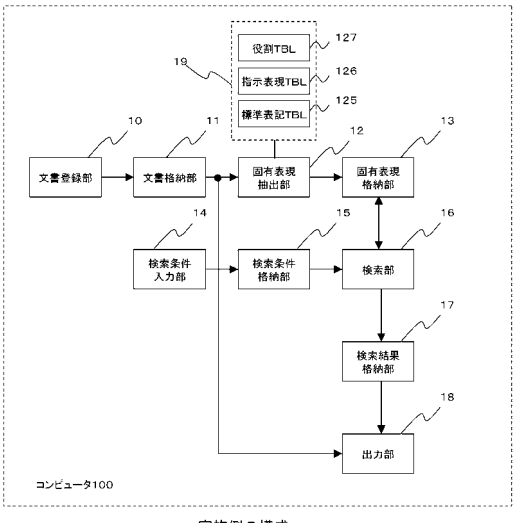
審査請求 未請求　請求項の数 9　OL　（全 16 頁）

(21)出願番号　特願2004-36053(P2004-36053)
(22)出願日　平成16年2月13日(2004.2.13)

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目１７番２２号
(74)代理人　100086531
弁理士　澤田　俊夫
(74)代理人　100093241
弁理士　宮田　正昭
(74)代理人　100101801
弁理士　山田　英治
(72)発明者　山下　明男
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
(72)発明者　永峯　猛志
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
最終頁に続く

(54)【発明の名称】文書検索装置および方法

(57)【要約】
【課題】　固有表現を参照する指示表現に着目して漏れのない検索・強調表示を行なう。
【解決手段】　固有表現抽出部１２は、文書中の固有表現および指示表現を抽出し、文書ＩＤ、固有表現（出現形）、固有表現（標準形）、カテゴリ、オフセット、長さ、役割を固有表現格納部１３に格納する。指示表現にはＰＲＯＮＯＵＮというカテゴリを付す。検索条件入力部１４は、「検索文字列」、「カテゴリ」、「指示表現検索フラグ」、「役割」、「文書ＩＤ」のうちの任意の１つを指定した検索条件を入力する。検索部１６は、検索条件格納部１５の検索条件を用いて、固有表現格納部１３に格納されている固有表現（指示表現を含む）のレコードを検索し、検索結果を検索結果格納部１７に格納する。出力部１８は、検索結果格納部１７の検索結果および文書格納部１１の文書内容を参照して文書を、固有表現および指示表現の検索結果部分を強調表示して、出力する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
文書を登録する文書登録手段と、
登録された文書を格納する文書格納手段と、
上記文書格納手段に格納されている文書を固有表現と指示表現とに基づいて検索する検索手段と、
上記検索手段による検索結果を出力する検索結果出力手段とを有することを特徴とする文書検索装置。
【請求項２】
固有表現および指示表現にはカテゴリが割り当てられ、文書中の指示表現の各々に対して、文書中の固有表現のうち、指示表現のカテゴリに対応するカテゴリを有し、文書中の位置が最も近いものを選択して、当該指示表現が指示する固有表現と判別する文書検索装置。
【請求項３】
所定の固有表現が検索されたときに、当該固有表現を指示する指示表現を検索結果に含める請求項２記載の文書検索装置。
【請求項４】
上記検索結果出力手段は、上記検索手段により検索された固有表現および指示表現を強調表示して提示する請求項１、２または３記載の文書検索装置。
【請求項５】
上記検索手段は、上記固有表現のカテゴリを指定して検索を行なう請求項１、２、３または４記載の文書検索装置。
【請求項６】
上記検索手段は、上記固有表現の文字列を指定して検索を行なう請求項１、２、３、４または５記載の文書検索装置。
【請求項７】
文書を入力する文書手段と、
上記文書入力手段により入力された文書を固有表現と指示表現とに基づいて検索する検索手段と、
上記検索手段による検索結果を出力する検索結果出力手段とを有することを特徴とする文書検索装置。
【請求項８】
文書登録手段が文書を登録するステップと、
文書格納手段が登録された文書を格納するステップと、
検索手段が上記文書格納手段に格納されている文書を固有表現と指示表現とに基づいて検索するステップと、
検索結果出力手段が上記検索手段による検索結果を出力するステップとを有することを特徴とする文書検索方法。
【請求項９】
文書登録手段が文書を登録するステップと、
文書格納手段が登録された文書を格納するステップと、
検索手段が上記文書格納手段に格納されている文書を固有表現と指示表現とに基づいて検索するステップと、
検索結果出力手段が上記検索手段による検索結果を出力するステップとをコンピュータに実行させるために用いられることを特徴とする文書検索用コンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は、重要な箇所を漏れなく提示する検索技術に関する。
【背景技術】

【０００２】
従来の全文検索は、指定された条件に適合する文字列を検索するものであり、例えば、検索条件中の文字列（単語）と適合する文書中の文字列を強調表示するものである（たとえば、Ｇｏｏｇｌｅ、商標、ＣｏｎｃｅｐｔＢａｓｅ　Ｓｅａｒｃｈ、商標）。ところで、文中において同一概念のものが再度表れる場合には、繰り返しを避けて、「同社」、「同製品」等の代替表現で置き換えられるため、本来、強調表示をして重要な箇所であることを表示すべきであるのに、文字列等に関する検索条件が適合しないため、検索漏れが生じ、この結果、ユーザが重要な箇所を見落とす原因となっている。
【０００３】
本発明者らは、このような代替表現（以下、指示表現という）も検索の対象とできるようにし、重要な箇所を見落とすことがないようにすることが重要であるという知見に到達した。
【０００４】
なおこの発明と関連する特許文献としては以下のものがある。
【０００５】
特許文献１は、組織名称と組織構成に関する情報を含む組織データと、氏名と所属組織に関する情報を含む人データを、その過去の変更履歴とともに人事ＤＢに記憶することを開示している。文書データは、その属性情報とともに文書ＤＢに記憶される。記憶された文書データを検索するための検索条件として、文書の属性情報が入力されると、入力された検索条件を人事ＤＢに記憶された変更履歴を基に有効な検索条件に補正し、あるいは限定する。過去の組織名で検索しても適切に検索を行なえる。しかしながら、この提案は、属性情報に基づいて検索を行うものであり、また、固有表現を抽出したり、その指示表現を考慮したりするものではない。
【０００６】
特許文献２は、テキストから単語インデックスを作成する際に、従来のように単語の見出しとその単語を含むテキストの情報だけでなく、その単語の品詞、属性の情報を追加登録することによって、品詞，属性（たとえば、固有表現）を検索条件とした検索を可能とし、特に多義性のある単語を検索キーとしたときの検索過剰を抑制する。また、単語の品詞や追加された属性に基づいて多義性のある語を展開してインデックスに登録し、多義性のある語を正しく展開した単語インデックスを生成する。これにより、検索キーと異表記でも同義の語を検索することができて検索漏れを低減でき、かつ、ユーザ自身が検索キーを同義語に展開して検索する場合に比べて検索負荷を抑えることができる。
【０００７】
しかしながら、この提案は固有表現を参照する指示表現に着目するものではない。
【０００８】
特許文献３は、データベースに登録すべき文書の記述から固有名詞を抽出し、抽出された固有名詞を登録すべき文書の付加情報として付加してデータベースに登録する。データベースに対して検索をするための入力を受け付けると、受け付けられた入力によりデータベースを検索し、検索され得られた文書の検索するための入力との適合度をスコアとして算出し、検索され得られた文書が有する付加情報が一致するごとに得られた文書についてスコアを加算して固有名詞スコアとして算出する。算出された固有名詞スコアはその固有名詞とともに表示される。
【０００９】
この提案も固有表現の指示表現に着目するものではない。
【特許文献１】特開平１０－２７１８０号公報
【特許文献２】特開平１１－３９３４７号公報
【特許文献３】特開２００１－２７３３２８公報
【発明の開示】
【発明が解決しようとする課題】
【００１０】

この発明は、以上の事情を考慮してなされたものであり、指示表現に着目して重要な箇所を漏れなくユーザに提示する検索技術を提供することを目的としている。

【課題を解決するための手段】

【００１１】

この発明の具体的の構成例では、上述の目的を達成するために、文書中に存在する固有表現を検索する際に、固有表現を参照している指示表現も検索の対象とし、重要な箇所を見逃すことが少なくなるような検索を実現する。文章内容の理解に必要な箇所の検出・表示を図り、読解を支援する。指示表現は、代名詞（「これ」、「それ」等）、代連体詞（「この」、「その」等）の指示詞のほか、広く、他の語句を指示、参照する語句をいい、「同社」、「同製品」等も含まれる。

【００１２】

さらにこの発明を説明する。

【００１３】

この発明の一側面によれば、上述の目的を達成するために、文書検索装置に：文書を登録する文書登録手段と；登録された文書を格納する文書格納手段と；上記文書格納手段に格納されている文書を固有表現と指示表現とに基づいて検索する検索手段と；上記検索手段による検索結果を出力する検索結果出力手段とを設けるようにしている。

【００１４】

この構成においては、固有表現のほかに指示表現にも着目して検索を行なうので重要な箇所を見落とすことが少なくなる。

【００１５】

固有表現（固有名ともいう）は、人名、組織名、地名、通貨、日付等、文中の重要な表現単位である。

【００１６】

この構成において、上記検索結果出力手段は、上記検索手段により検索された固有表現や指示表現を強調表示して提示する。該当する固有表現や指示表現自体を強調表示しても良いし、該当する固有表現や指示表現を含む文章単位やパラグラフ単位を強調表示しても良い。該当する固有表現や指示表現を含む文書全体を表示するのでなく、その要約を表示したり、該当する文章やパラグラフを表示するのでもよい。要約、該当文章、該当パラグラフを表示する際に該当する固有表現や指示表現を強調表示しても良い。

【００１７】

また、例えば、固有表現および指示表現にはカテゴリが割り当てられ、文書中の指示表現の各々に対して、文書中の固有表現のうち、指示表現のカテゴリに対応するカテゴリを有し、文書中の位置が最も近いものを選択して、当該指示表現が指示する固有表現と判別するようにしてもよい。この場合、所定の固有表現が検索されたときに、当該固有表現を指示する指示表現を検索結果に含める。

【００１８】

また、上記検索手段は、上記固有表現のカテゴリを指定して検索を行なうようにしてもよい。カテゴリは例えば組織、人名、地名等であるが、さらにブレークダウンしてもよい。

【００１９】

また、上記検索手段は、上記固有表現の文字列を指定して検索を行なうようにしてもよい。

【００２０】

また、この発明の他の側面によれば、上述の目的を達成するために、文書検索装置に：文書を入力する文書手段と；上記文書入力手段により入力された文書を固有表現と指示表現とに基づいて検索する検索手段と；上記検索手段による検索結果を出力する検索結果出力手段とを設けるようにしている。

【００２１】

この構成においても、固有表現のほかに指示表現にも着目して検索を行なうので重要な

箇所を見落とすことが少なくなる。

【００２２】

なお、この発明は装置またはシステムとして実現できるのみでなく、方法としても実現可能である。また、そのような発明の一部をソフトウェアとして構成することができることはもちろんである。またそのようなソフトウェアをコンピュータに実行させるために用いるソフトウェア製品もこの発明の技術的な範囲に含まれることも当然である。

【００２３】

この発明の上述の側面および他の側面は特許請求の範囲に記載され以下実施例を用いて詳述される。

【発明の効果】

【００２４】

この発明によれば、固有表現のほかに指示表現にも着目して検索を行なうので重要な箇所を見落とすことが少なくなる。

【発明を実施するための最良の形態】

【００２５】

以下、この発明の実施例について説明する。

【００２６】

図１は、この発明を適用した文書検索装置の実施例を全体として示しており、この図において、文書検索装置は、文書登録部１０、文書格納部１１、固有表現抽出部１２、固有表現格納部１３、検索条件入力部１４、検索条件格納部１５、検索部１６、検索結果格納部１７および出力部１８等を含んで構成されている。図に示す各機能ブロックの一部は実際にはコンピュータ１００にコンピュータプログラムとして実装される。

【００２７】

文書登録部１０は、１または複数の文書を入力するものである。ユーザにより指定された文書を入力しても良いし、送出先が指定した文書をそのまま受け取って入力しても良いし、あるいは、キーワードや単語ベクトル等を用いて文書を分類して所定の分類グループの文書を入力しても良い。文書登録部１０により入力された文書は文書格納部１１に格納される。格納された文書は例えば図８に示すようなものであり、文書ＩＤが振られる。なお、事例の文等に含まれる会社名、商品名はいずれも商標であり、また人名は実在の人物を表示するものではなく架空のものである。文書は、文書属性と文書コンテンツと個別に管理しても良い。

【００２８】

固有表現抽出部１２は、文書中の固有表現（固有名ともいう）を抽出するものである。固有表現は、人名、組織名、地名、通貨、日付等、文中の重要な表現単位である。固有表現抽出部１２は、例えば、図２に示すように、形態素解析部１２１、形態素解析辞書記憶部１２２、ルール適用部１２３およびルール記憶部１２４等を含んで構成される。

【００２９】

固有表現抽出部１２の入力は例えば図３に示すようなものであり、形態素解析辞書記憶部１２２の形態素解析辞書は例えば図４に示すようなエントリを持つ。形態素解析部１２１は形態素解析辞書を用いて入力例（図３）から図５に示すような解析結果を得る。図５において「／」は形態素間の区切りを示し、「＜」、「＞」で囲む部分は品詞を表す。図では、開始位置や長さは省略している。形態素解析結果はルール適用部１２３に入力されてルール記憶部１２４の抽出ルールを参照して固有表現が抽出される。図６は抽出ルールの例を示し、例えばルール番号５により「姓」と「名」が結合されて「ＰＥＲＳＯＮ」のカテゴリが付される。抽出結果は図７に示すようになる。この例では、各固有表現が抽出され、＜ＯＲＧＡＮＩＺＡＴＩＯＮ＞、＜ＰＥＲＳＯＮ＞、＜ＣＵＲＲＥＮＣＹ＞、＜ＤＡＴＥ＞、＜ＰＬＡＣＥ＞等のカテゴリが付される。

【００３０】

さらに固有表現および指示表現の抽出処理について説明する。図９は、固有表現抽出部１２の処理を示しており、この図において、対象となる個々の文書を順次に取り出し、文

書内容（図８に示す）に対して固有表現抽出を行う（Ｓ１０）。抽出された固有表現の情報を固有表現格納部１３に格納する（Ｓ１１）。固有表現の情報は、例えば、文書ＩＤ、固有表現（出現形）、固有表現（標準形）、カテゴリ、オフセット、長さであるが、これに限定されない。固有表現（標準形）は、図１０の標準標記テーブル１２５を検索して決めることができる。図１２に示す役割テーブル（役割・助詞（相当語句）対応表）」１２７を参照し、役割に関する助詞（相当語句）が、抽出した固有表現に続いていれば、その固有表現の役割として対応する役割も登録する（Ｓ１２）。図１１に示す指示表現テーブル１２６の指示表現が抽出されたら、これも固有表現格納部１３に格納する（Ｓ１３）。そのカテゴリはＰＲＯＮＯＵＮとする。指示表現テーブルで指定されたカテゴリを持つ固有表現を固有表現格納部１３から検索し、もっとも近い固有表現のレコード番号を参照先として登録する。以上の処理を文書単位に実行する（Ｓ１４）。標準表記テーブル１２５（図１０）、指示表現テーブル１２６（図１１）、役割テーブル１２７（図１２）はコンピュータ１００の記憶部１９に記憶される。

【００３１】

なお、先に述べたように指示表現は指示表現テーブル１２６（図１１）を用いて表引きないし辞書引きにより判別されるが、指示表現抽出用のルールを設け、例えば、「同」、「上記」、「この」等を含む一連の語句（複合語を含む）から指示表現を導出するようにしても良い。この場合、一連の語句の主要な語句例えば「この製品」の場合には「製品」のカテゴリを指示表現のカテゴリに割り当てることができる。

【００３２】

抽出された固有表現（指示表現を含む）は図１３に示すように固有表現格納部１３に格納される。この例では、抽出された固有表現に対して、レコード番号、所属する文書ＩＤ、固有表現（出現形）、固有表現（標準形）、カテゴリ、オフセット、長さ、役割、参照先等が与えられる。

【００３３】

図１に戻る。検索条件入力部１４は例えば所定のグラフィカルユーザインタフェース（ＧＵＩ）を用いてユーザにより入力される。管理者等により、予め固定した検索条件が指定される場合もある。以下の例では、検索条件入力部１４から、「検索文字列」、「カテゴリ」、「指示表現検索フラグ」（指示表現を用いて検索するかどうかの指定）、「役割」、「文書ＩＤ」のうちの任意の１つを指定した検索条件を入力する（図１６～図２４）。「文書ＩＤ」は、文書名でもよいし、他の文書属性を用いた検索でも良い。検索条件は検索条件格納部１５に格納される。検索部１６は、検索条件格納部１５の検索条件を用いて、固有表現格納部１３に格納されている固有表現のレコードを検索し、検索結果を検索結果格納部１７に格納する。

【００３４】

出力部１８は、検索結果格納部１７の検索結果および文書格納部１１の文書内容を参照して文書を、検索結果部分（検索された固有表現部分）を強調表示して、出力する（図１６～図２４参照）。もちろん、該当する文書の全文でなく、要約を自動生成し、これに検索された固有表現部分を強調表示して出力するようにしても良い。

【００３５】

さらに図１４を参照して検索部１６の処理を説明する。図１４において、検索条件入力部１４は、ユーザが例えばＧＵＩを通して設定した検索条件を、検索条件格納部１５に登録する（Ｓ２０）。検索部１６は、検索条件格納部１５の検索文字列、カテゴリ、文書ＩＤに適合する、文書あるいは文書内での出現箇所を検索し、その結果を検索結果格納部１７に、そのレコード番号と表示フラグ（登録時はＯＮ）を図１５に示すように格納する（Ｓ２１）。なお、検索条件格納部１５の検索条件に文書ＩＤが指定されていない場合には固有表現格納部１３の全レコードに対して検索を実行する。文書ＩＤが指定されていた場合には、その文書ＩＤのレコード群に対して検索を実行する。また、検索条件格納部１５の検索条件で指示表現検索フラグがＯＮであれば、検索された固有表現のレコード番号を持ち、カテゴリがＰＲＯＮＯＵＮで文書ＩＤが同じである指示表現のレコードも検索して

結果を検索結果格納部１７に登録する。検索条件格納部１５の検索条件で役割が指定されていれば、検索された固有表現のレコードの中で更に、指定された役割でないレコードの表示フラグをＯＦＦにする（先に述べたように表示フラグはＯＮとして登録されている）。検索結果格納部１７の検索結果レコードの表示フラグのＯＮ／ＯＦＦを切り替えて役割指定した場合の検索結果としない場合の検索結果とを即座に比較することもできる。

【００３６】

さらに検索例を用いて検索処理を説明する。

【００３７】

［カテゴリを指定した検索（指示表現検索フラグはＯＦＦ）］

図１６（ａ）は、特定のカテゴリ（この例では「ＯＲＧＡＮＩＺＡＴＩＯＮ」）の固有表現を検索条件とするものである。指示表現検索フラグはＯＦＦであり、該当する語句が検索・強調表示され、他方、指示表現については検索・強調表示されない。検索結果は図１６（ｂ）に示すようになり、役割を指定していないので、表示フラグは当初の「ＯＮ」のままであり、レコード番号１（文書ＩＤ＝１、図１３参照）、レコード番号６（文書ＩＤ＝２）、レコード番号７（文書ＩＤ＝２）、レコード番号８（文書ＩＤ＝３）、レコード番号９（文書ＩＤ＝３）、レコード番号１０（文書ＩＤ＝３）、レコード番号１３（文書ＩＤ＝３）が検索され、かつ表示される。図１６（ｃ）に示すように、組織名のカテゴリの固有表現を強調表示して文書が表示される。図では強調部分に下線を付したが、色を変える、ブリンク属性を付す等種々採用できる。矢印等のマークを付しても良い。

【００３８】

［カテゴリと役割を指定した検索（指示表現検索フラグはＯＦＦ）］

図１７の例は、特定のカテゴリに加えて役割を検索条件に含めるものである。この例では、カテゴリが「ＯＲＧＡＮＩＺＡＴＩＯＮ」であり（図１７（ａ））、役割が「主体」である。この場合も、指示表現検索フラグはＯＦＦであり、該当する語句が検索・強調表示され、他方、指示表現については検索・強調表示されない。図１６の場合と同様に、レコード番号１、６、７、８、９、１０、１３が検索結果に含まれる。ただし、役割が「主体」と指定されているので、それ以外の役割のレコード番号７、１０、１３は表示フラグが「ＯＦＦ」にリセットされる（図１７（ｂ））。したがって、図１７（ｃ）の出力例では、図１６（ｃ）に較べて一部の固有表現の強調表示がリセットされる。

【００３９】

［文字列、役割および文書ＩＤを指定した検索（指示表現検索フラグはＯＦＦ）］

図１８の例は、検索条件に「富士ゼロックス」（商標）という文字列（キーワード）、「主体」という役割、文書ＩＤ＝２を含ませている（図１８（ａ））。この場合も、指示表現検索フラグはＯＦＦであり、該当する語句が検索・強調表示され、他方、指示表現については検索・強調表示されない。検索結果は図１８（ｂ）に示すとおりであり、出力結果は図１８（ｃ）に示すとおりである。

【００４０】

［文字列を指定した検索（指示表現検索フラグはＯＦＦ）］

図１９の例は、検索条件に「富士ゼロックス」（商標）という文字列（キーワード）を含ませている（図１９（ａ））。この場合も、指示表現検索フラグはＯＦＦであり、該当する語句が検索・強調表示され、他方、指示表現については検索・強調表示されない。検索結果は図１９（ｂ）に示すとおりであり、出力結果は図１９（ｃ）に示すとおりである（この例では「富士ゼロックス株式会社」または「富士ゼロックス」（出現形）は商標）。

【００４１】

［文字列を指定した検索（指示表現検索フラグはＯＮ）］

図２０の例は、検索条件に「富士ゼロックス」（商標）という文字列（キーワード）を含ませている（図２０（ａ））。この場合は、指示表現検索フラグはＯＮであり、該当する語句が検索・強調表示されるとともに、これら語句を参照する指示表現についても検索・強調表示される。検索結果は図２０（ｂ）に示すとおりであり、出力結果は図２０（ｃ

）に示すとおりである。指示表現検索フラグをＯＮにすることにより、図１９の例に較べて、文書１の指示表現の「同社」および「このメーカ」が検索・強調表示されている。

【００４２】

［文字列および文書ＩＤを指定した検索（指示表現検索フラグはＯＦＦ）］

図２１の例は、検索条件に「富士ゼロックス」（商標）という文字列（キーワード）および文書ＩＤ＝１を含ませている（図２１（ａ））。指示表現検索フラグはＯＦＦであり、該当する語句が検索・強調表示され、他方、指示表現については検索・強調表示されない。検索結果は図２１（ｂ）に示すとおりであり、出力結果は図２１（ｃ）に示すとおりである。当該指定された文書が、当該指定文字列（出現形）を強調表示されて表示される。

【００４３】

［文字列および文書ＩＤを指定した検索（指示表現検索フラグはＯＮ）］

図２２の例は、検索条件に「富士ゼロックス」（商標）という文字列（キーワード）および文書ＩＤ＝１を含ませている（図２２（ａ））。図２１の例に対して、この例では、指示表現検索フラグはＯＮであり、該当する語句が検索・強調表示されるとともに、これら語句を参照する指示表現についても検索・強調表示される。検索結果は図２２（ｂ）に示すとおりであり、出力結果は図２２（ｃ）に示すとおりである。指示表現検索フラグをＯＮにすることにより、指示表現の「同社」および「このメーカ」が検索・強調表示されている。

【００４４】

［すべてのカテゴリを指定した検索（指示表現検索フラグはＯＦＦ）］

図２３の例は、検索条件を、すべてのカテゴリにしたものである（図２３（ａ）参照）。この場合、図２３（ｂ）に示すように、すべての固有表現が強調表示される。ただし、指示表現が強調表示されない。

【００４５】

［カテゴリおよび文書ＩＤを指定した検索（指示表現検索フラグはＯＦＦ）］

図２４の例は、検索条件にカテゴリおよび文書ＩＤを指定したものである（図２４（ａ）参照）。図２４（ｂ）に示すように、当該指定された文書が、当該指定カテゴリの固有表現を強調表示されて表示される。ただし、指示表現は強調表示されない。

【００４６】

なお、この発明は上述の実施例に限定されるものではなくその趣旨を逸脱しない範囲で種々変更が可能である。例えば、上述の例では、コンピュータ１００に文書検索装置を実装したが、図２５や図２６に示すように文書検索装置をクライアントサーバの形態で実装しても良い。図２５の例では、検索条件入力部１４および出力部１８がクライアントコンピュータ２００に配置され、残りがサーバコンピュータ３００に配置されている。また図２６の例では、文書登録部１０もクライアントコンピュータ２００に配置されている。また、図２７に示すように、ユーザが文書入力部２０を介して文書を指定して入力して即座に固有名抽出を行なって出力するようにしても良い。例えば、所定の文書処理アプリケーションのプラグインとして文書検索装置のプログラムを実装して、処理中の文書の固有名抽出を行なうようにしても良い。

【図面の簡単な説明】

【００４７】

【図１】この発明の実施例を全体として示すブロック図である。

【図２】上述実施例の固有表現抽出部の構成例を説明するブロック図である。

【図３】上述実施例の固有表現抽出部に入力される文書の例を説明する図である。

【図４】上述実施例の固有表現抽出部で用いる形態素解析辞書の例を説明する図である。

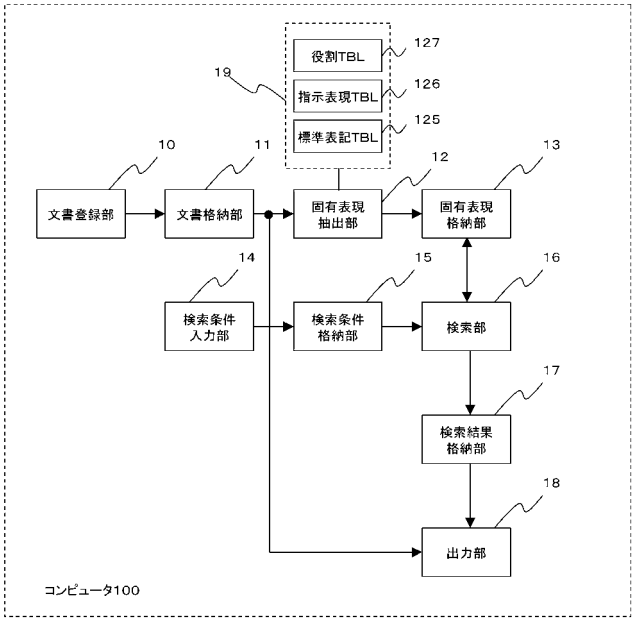
【図５】上述実施例の固有表現抽出部における形態素解析結果の例を説明する図である。

【図６】上述実施例の固有表現抽出部による固有名抽出ルールの説明する図である。

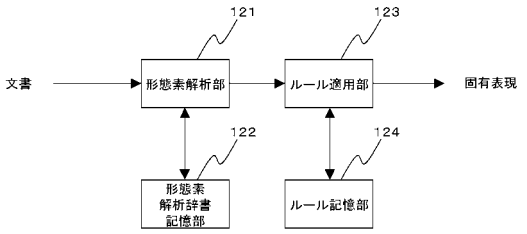
【図７】上述実施例の固有表現抽出部による固有名抽出結果を簡略化して説明する図である。

【図８】上述実施例の入力文書の例を説明する図である。
【図９】上述実施例の固有表現抽出部の動作を説明するフローチャートである。
【図１０】上述実施例の固有表現抽出部で用いる標準表記テーブルを説明する図である。
【図１１】上述実施例の固有表現抽出部で用いる指示表現テーブルを説明する図である。
【図１２】上述実施例の固有表現抽出部で用いる役割テーブル（役割・助詞対応）を説明する図である。
【図１３】上述実施例の固有表現抽出部により抽出された固有表現レコードを説明する図である。
【図１４】上述実施例の検索部の動作を説明するフローチャートである。
【図１５】上述検索部で検索されたレコードを説明する図である。
【図１６】上述検索部による検索例を説明する図である。
【図１７】上述検索部による他の検索例を説明する図である。
【図１８】上述検索部による他の検索例を説明する図である。
【図１９】上述検索部による他の検索例を説明する図である。
【図２０】上述検索部による他の検索例を説明する図である。
【図２１】上述検索部による他の検索例を説明する図である。
【図２２】上述検索部による他の検索例を説明する図である。
【図２３】上述検索部による他の検索例を説明する図である。
【図２４】上述検索部による他の検索例を説明する図である。
【図２５】上述実施例の変形例を説明するブロック図である。
【図２６】上述実施例の他の変形例を説明するブロック図である。
【図２７】上述実施例の他の変形例を説明するブロック図である。
【符号の説明】
【００４８】
１０　文書登録部
１１　文書格納部
１２　固有表現抽出部
１３　固有表現格納部
１４　検索条件入力部
１５　検索条件格納部
１６　検索部
１７　検索結果格納部
１８　出力部
１００　コンピュータ
１２１　形態素解析部
１２２　形態素解析辞書記憶部
１２３　ルール適用部
１２４　ルール記憶部
２００　クライアントコンピュータ
３００　サーバコンピュータ

【図１】



【図２】



固有表現抽出部の構成例

【図３】

富士ゼロックス
山田、太郎、山田太郎
１００ドル、12000円
2002年03月11日
平成１４年３月１１日
ホゲホゲ株式会社
東京都港区赤坂

入力文書の例

（＊）会社名、商品名は各社の商標である。人物名は架空のものである

【図４】

ドル	通貨単位	その他情報
：		
円	通貨単位	
株式会社	接尾辞_組織	
区	接尾辞_地名	
月	助数詞-月	
港	地名	
山田	姓	
赤坂	地名	
太郎	名	
都	接尾辞_地名	
東京	地名	
日	助数詞-日	
年	助数詞-年	
富士ゼロックス	組織	
平成	年号	

形態素解析辞書の例

（＊）会社名、商品名は各社の商標である。人物名は架空のものである

【図５】

／富士ゼロックス<組織>／
／山田<姓>／、<記号>／太郎<名>／、<記号>／山田<姓>／太郎<名>／
／１００<数>／ドル<通貨単位>／、<記号>／12000<数>／円<通貨単位>／
／2002<数>／年<助数詞-年>／03<数>／月<助数詞-月>／11<数>／日<助数詞-日>／
／平成<年号>／１４<数>／年<助数詞-年>／３<数>／月<助数詞-月>／１１<数>／日<助数詞-日>／
／ホゲホゲ<未知語>／株式会社<接尾辞_組織>／
／東京<地名>／都<接尾辞_地名>／港<地名>／区<接尾辞_地名>／赤坂<地名>／

（＊）会社名、商品名は各社の商標である。人物名は架空のものである

形態素解析結果の例

【図６】

No	条件(品詞列のパターン)	処理	カテゴリ
1	組織		ORGANIZATION
2	未知語－接尾辞_組織	連結	ORGANIZATION
3	姓		PERSON
4	名		PERSON
5	姓＋名	連結	PERSON
6	数＋通貨単位	連結	CURRENCY
7	(年号)*－数＋助数詞-年＋数＋助数詞-月＋数＋助数詞-日	連結	DATE
8	地名		PLACE
9	地名+接尾辞_地名	連結	PLACE

固有表現抽出ルールの例

【図７】

／富士ゼロックス< ORGANIZATION >／
／山田< PERSON >／、／太郎< PERSON >／、／山田太郎< PERSON >／
／１００ドル< CURRENCY>／、／12000円< CURRENCY>／
／2002年03月11日< DATE>／
／平成１４年３月１１日< DATE>／
／ホゲホゲ株式会社< ORGANIZATION >／
／東京都<PLACE>／港区<PLACE>／赤坂<PLACE>／

（＊）会社名、商品名は各社の商標である。人物名は架空のものである

固有表現抽出結果の例

【図８】

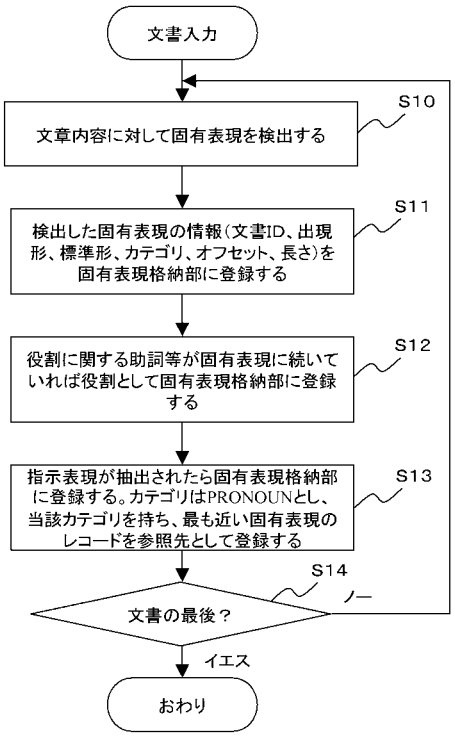
文書id	文書内容
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）
2	電子行政（電子政府・電子自治体）推進のために、富士ゼロックスはお客様のニーズにお応えするソリューションを提供していきます。 ○富士ゼロックスの「電子行政への取り組み」内容
3	セブン-イレブン・ジャパンと富士ゼロックスは、パソコンで作成した文書やデジカメで撮影した画像などを、セブン-イレブン店頭に設置したマルチコピー機からプリントアウトできる個人向けサービス「ネットプリントサービスbasic（ベーシック）」を2003年11月5日開始する。富士ゼロックスでは、「basic」サービスの開始を記念して、スタートキャンペーンを2004年1月8日まで実施する。

（＊）会社名、商品名は各社の商標である。

具体的な入力文書の例

【図９】

固有表現抽出部の処理



【図１０】

標準標記テーブルの例

固有表現（出現形）	固有表現（標準形）
富士ゼロックス株式会社	富士ゼロックス株式会社
富士ゼロックス	
Fuji Xerox	
FX	
セブン-イレブン・ジャパン	株式会社セブン-イレブン・ジャパン
セブン-イレブン	
セブンイレブン	
ネットプリントサービスbasic	ネットプリントサービスbasic
Net Print Service basic	

（＊）会社名、商品名は各社の商標である。

【図１１】

指示表現テーブルの例

カテゴリ	指示表現
ORGANIZATION	同社
	同校
	この会社
	このメーカー
PRODUCT	同製品
	この製品

【図１２】

役割テーブル（役割と助詞（相当語句）との対応）の例

役割	助詞（相当語句）の例
主体	が、は、も
対象	を、に に関して、について、に対して
その他	へ、で、と、から、より、まで、では、でも、 において、によって、のために、にとって、を用いて

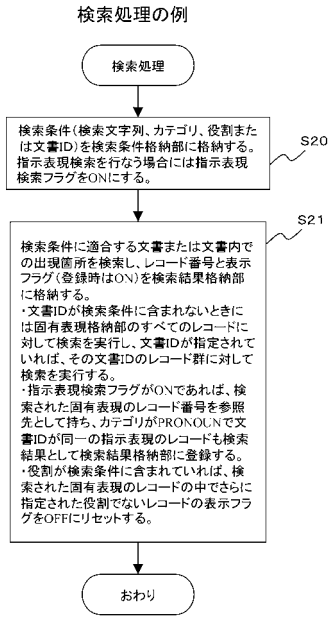
【図１３】

固有表現抽出結果の例

Rec No	文書ID	固有表現（出現形）	固有表現（標準形）	カテゴリ	オフセット（Byte）	長さ（文字）	役割	参照先
1	1	富士ゼロックス株式会社	富士ゼロックス株式会社	ORGANIZATION	0	11	主体(動作や状態の)	
2	1	9月30日	09-30	DATE	24	5		
3	1	DocuCentre	DocuCentre	PRODUCT	49	10		
4	1	同社	同社	PRONOUN	109	2		1
5	1	このメーカー	このメーカー	PRONOUN	160	6		1
6	2	富士ゼロックス	富士ゼロックス株式会社	ORGANIZATION	46	7	主体(動作や状態の	
7	2	富士ゼロックス	富士ゼロックス株式会社	ORGANIZATION	126	7	その他	
8	3	セブン-イレブン・ジャパン	株式会社セブン-イレブン・ジャパン	ORGANIZATION	0	13	主体(動作や状態の)	
9	3	富士ゼロックス	富士ゼロックス株式会社	ORGANIZATION	27	7	主体(動作や状態の)	
10	3	セブン-イレブン	株式会社セブン-イレブン・ジャパン	ORGANIZATION	99	8	その他	
11	3	2003年11月5日	2003-11-05	DATE	229	10		
12	3	ネットプリントサービスbasic	ネットプリントサービスbasic	PRODUCT	184	16		
13	3	富士ゼロックス	富士ゼロックス株式会社	ORGANIZATION	254	7	その他	
14	3	2004年1月8日	2004-01-08	DATE	331	10		

（＊）会社名、商品名は各社の商標である。

【図１４】



【図１５】

RecNo	
表示フラグ	

検索レコードの例

【図１６】

特定のカテゴリの固有表現を検索：役割は未指定：：指示表現検索フラグがOFF

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
	ORGANIZATION	OFF		

検索結果格納部

(b)

RecNo	1	6	7	8	9	10	13
表示フラグ	ON	ON	ON	ON	ON	ON	ON

(c)

文書id	文書内容（ヒットする部分を強調表示）
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）
2	電子行政（電子政府・電子自治体）推進のために、富士ゼロックスはお客様のニーズにお応えするソリューションを提供していきます。 ○富士ゼロックスの「電子行政への取り組み」内容
3	セブン-イレブン・ジャパンと富士ゼロックスは、パソコンで作成した文書やデジカメで撮影した画像などを、セブン-イレブン店頭に設置したマルチコピー機からプリントアウトできる個人向けサービス「ネットプリントサービスbasic（ベーシック）」を2003年11月5日開始する。 富士ゼロックスでは、「basic」サービスの開始を記念して、スタートキャンペーンを2004年1月8日まで実施する。

役割が指定されていないときは、カテゴリがORGANIZATIONのものが検索される（フィルタリングは行われない）。

（＊）会社名、商品名は各社の商標である。

【図１７】

特定のカテゴリの固有表現を検索：役割に主体を指定：指示表現検索フラグがOFF

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
	ORGANIZATION	OFF	主体	

検索結果格納部

(b)

RecNo	1	6	7	8	9	10	13
表示フラグ	ON	ON	OFF	ON	ON	OFF	OFF

(c)

文書id	文書内容（ヒットする部分を強調表示）
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）
2	電子行政（電子政府・電子自治体）推進のために、富士ゼロックスはお客様のニーズにお応えするソリューションを提供していきます。 ○富士ゼロックスの「電子行政への取り組み」内容
3	セブン-イレブン・ジャパンと富士ゼロックスは、パソコンで作成した文書やデジカメで撮影した画像などを、セブン-イレブン店頭に設置したマルチコピー機からプリントアウトできる個人向けサービス「ネットプリントサービスbasic（ベーシック）」を2003年11月5日開始する。 富士ゼロックスでは、「basic」サービスの開始を記念して、スタートキャンペーンを2004年1月8日まで実施する。

役割に主体を指定すると、カテゴリがORGANIZATIONで役割が主体のものが検索される

（＊）会社名、商品名は各社の商標である。

【図１８】

特定の固有表現を検索：役割に主体を指定：：指示表現検索フラグがOFF

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
富士ゼロックス		OFF	主体	2

検索結果格納部

(b)

RecNo	6	7
表示フラグ	ON	OFF

(c)

文書id	文書内容（ヒットする部分を強調表示）
2	電子行政（電子政府・電子自治体）推進のために、富士ゼロックスはお客様のニーズにお応えするソリューションを提供していきます。 ○富士ゼロックスの「電子行政への取り組み」内容

指示表現検索フラグがONのときは、指示表現も含めて検索する。

（＊）会社名、商品名は各社の商標である。

【図１９】

特定の固有表現を検索（本発明）：指示表現検索フラグがOFF

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
富士ゼロックス		OFF		

検索結果格納部

(b)

RecNo	1	6	7	9	13
表示フラグ	On	On	On	On	On

(c)

文書id	文書内容（ヒットする部分を強調表示）
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）
2	電子行政（電子政府・電子自治体）推進のために、富士ゼロックスはお客様のニーズにお応えするソリューションを提供していきます。 ○富士ゼロックスの「電子行政への取り組み」内容
3	セブン-イレブン・ジャパンと富士ゼロックスは、パソコンで作成した文書やデジカメで撮影した画像などを、セブン-イレブン店頭に設置したマルチコピー機からプリントアウトできる個人向けサービス「ネットプリントサービスbasic（ベーシック）」を2003年11月5日開始する。 富士ゼロックスでは、「basic」サービスの開始を記念して、スタートキャンペーンを2004年1月8日まで実施する。

指示表現検索フラグがOFFのときは、指示表現は検索しない。

（＊）会社名、商品名は各社の商標である。

【図２０】

特定の固有表現を検索（本発明）：指示表現検索フラグがON

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
富士ゼロックス		ON		

検索結果格納部

(b)

RecNo	1	4	5	6	7	9	13
表示フラグ	On	On	On	On	On	On	On

(c)

文書id	文書内容（ヒットする部分を強調表示）
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）
2	電子行政（電子政府・電子自治体）推進のために、富士ゼロックスはお客様のニーズにお応えするソリューションを提供していきます。 ○富士ゼロックスの「電子行政への取り組み」内容
3	セブン-イレブン・ジャパンと富士ゼロックスは、パソコンで作成した文書やデジカメで撮影した画像などを、セブン-イレブン店頭に設置したマルチコピー機からプリントアウトできる個人向けサービス「ネットプリントサービスbasic（ベーシック）」を2003年11月5日開始する。 富士ゼロックスでは、「basic」サービスの開始を記念して、スタートキャンペーンを2004年1月8日まで実施する。

指示表現検索フラグがOFFのときは、指示表現も検索する。

（＊）会社名、商品名は各社の商標である。

【図２１】

特定文書内で、特定の固有表現を検索

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
富士ゼロックス		OFF		1

検索結果格納部

(b)

RecNo	1
表示フラグ	On

(c)

文書id	文書内容（ヒットする部分を強調表示）
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）

指示表現検索フラグがOFFのときは、指示表現は検索しない。

（＊）会社名、商品名は各社の商標である。

【図２２】

特定の固有表現を検索（本発明）：指示表現検索フラグがON

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
富士ゼロックス		ON		1

(b)

検索結果格納部

RecNo	1	4	5
表示フラグ	On	On	On

(c)

文書id	文書内容（ヒットする部分を強調表示）
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）

指示表現検索フラグがONのときは、指示表現も含めて検索する。

（＊）会社名、商品名は各社の商標である。

【図２３】

すべての固有表現を検索（従来と同様）

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
	ALL	OFF		

(b)

文書id	文書内容（ヒットする部分を強調表示）
1	富士ゼロックス株式会社は9月30日、デジタル複合機「DocuCentre」シリーズに新ラインアップを追加すると発表した。同社では、年間20,000台の販売を目標としている。（注：このメーカーの最新のリリース記事は…）
2	電子行政（電子政府・電子自治体）推進のために、富士ゼロックスはお客様のニーズにお応えするソリューションを提供していきます。 ○富士ゼロックスの「電子行政への取り組み」内容
3	セブン-イレブン・ジャパンと富士ゼロックスは、パソコンで作成した文書やデジカメで撮影した画像などを、セブン-イレブン店頭に設置したマルチコピー機からプリントアウトできる個人向けサービス「ネットプリントサービスbasic（ベーシック）」を2003年11月5日開始する。 富士ゼロックスでは、「basic」サービスの開始を記念して、スタートキャンペーンを2004年1月8日まで実施する。

文書idが空白であり、すべての文書を検索対象とする。
固有表現格納部に登録されている固有表現の出現箇所をすべて検索する（固有名抽出の結果、富士ゼロックス株式会社も強調表示可能）

（＊）会社名、商品名は各社の商標である。

【図２４】

特定文書内でカテゴリがORGANIZATIONの固有表現を検索

(a)

検索文字列	カテゴリ	指示表現検索フラグ	役割	文書id
	ORGANIZATION	OFF		3

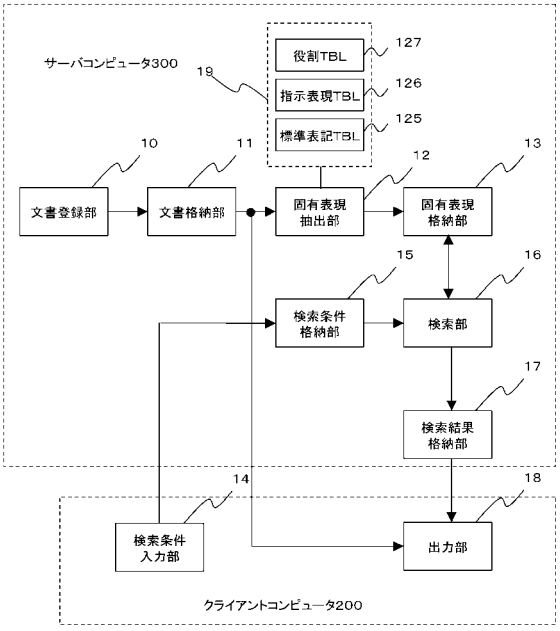
(b)

文書id	文書内容（ヒットする部分を強調表示）
3	セブン-イレブン・ジャパンと富士ゼロックスは、パソコンで作成した文書やデジカメで撮影した画像などを、セブン-イレブン店頭に設置したマルチコピー機からプリントアウトできる個人向けサービス「ネットプリントサービスbasic（ベーシック）」を2003年11月5日開始する。 富士ゼロックスでは、「basic」サービスの開始を記念して、スタートキャンペーンを2004年1月8日まで実施する。

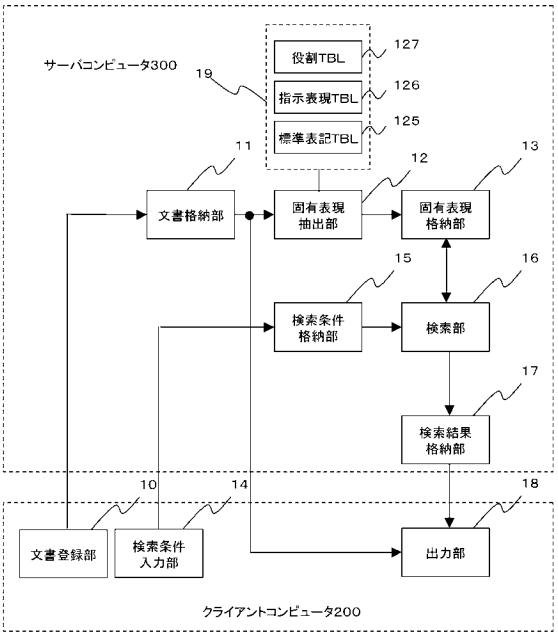
文書id=3の内容に対して、カテゴリが「ORGANIZATION」である箇所を固有表現格納部から検索する。

（＊）会社名、商品名は各社の商標である。

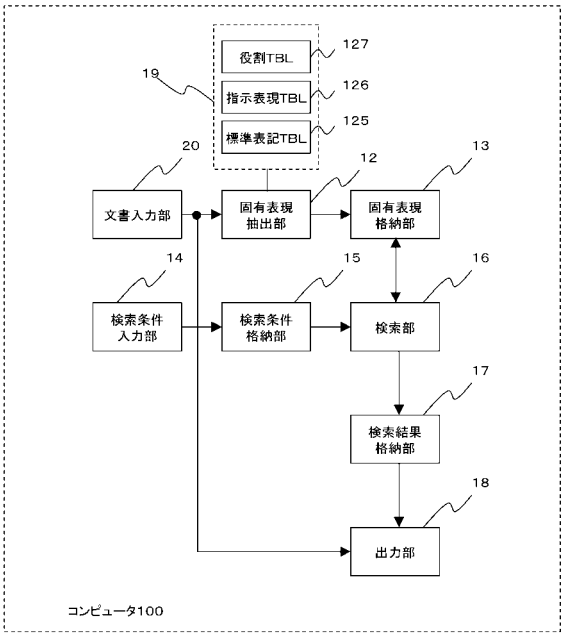
【図２５】



【図２６】


サーバコンピュータ300
19
役割TBL
127
指示表現TBL
126
標準表記TBL
125
11
文書格納部
12
固有表現
抽出部
13
固有表現
格納部
15
検索条件
格納部
16
検索部
17
検索結果
格納部
10
文書登録部
14
検索条件
入力部
18
出力部
クライアントコンピュータ200

【図２７】


19
役割TBL
127
指示表現TBL
126
標準表記TBL
125
20
文書入力部
12
固有表現
抽出部
13
固有表現
格納部
14
検索条件
入力部
15
検索条件
格納部
16
検索部
17
検索結果
格納部
18
出力部
コンピュータ100

フロントページの続き

(72)発明者　芳地　克典
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
(72)発明者　梅基　宏
神奈川県川崎市高津区坂戸３丁目２番１号　ＫＳＰ　Ｒ＆Ｄ　ビジネスパークビル　富士ゼロックス株式会社内

Ｆターム(参考) 5B009 QA03 RB32 SA03 VA02
5B075 ND03 NK32 NK35 NR06 NR12 PQ02 PQ22 QP01 QP03 UU06