

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4368550号
(P4368550)

(45) 発行日 平成21年11月18日(2009.11.18)

(24) 登録日 平成21年9月4日(2009.9.4)

(51) Int.Cl. F I
G06F 17/30 (2006.01)
 G06F 17/30 330C
 G06F 17/30 170A
 G06F 17/30 340Z

請求項の数 7 (全 15 頁)

(21) 出願番号	特願2001-401817 (P2001-401817)	(73) 特許権者	390024350 株式会社ジャストシステム
(22) 出願日	平成13年12月28日 (2001.12.28)		徳島県徳島市川内町平石若松108番地4
(65) 公開番号	特開2003-196309 (P2003-196309A)	(74) 代理人	100104190 弁理士 酒井 昭徳
(43) 公開日	平成15年7月11日 (2003.7.11)	(72) 発明者	出口 知哲 徳島市沖浜東3丁目46番地 株式会社ジャストシステム内
審査請求日	平成16年10月12日 (2004.10.12)	(72) 発明者	平本 真一 徳島市沖浜東3丁目46番地 株式会社ジャストシステム内
審判番号	不服2007-15060 (P2007-15060/J1)	(72) 発明者	菊地 文子 徳島市沖浜東3丁目46番地 株式会社ジャストシステム内
審判請求日	平成19年5月24日 (2007.5.24)		

最終頁に続く

(54) 【発明の名称】 文書検索装置、文書検索方法およびその方法をコンピュータに実行させるプログラム

(57) 【特許請求の範囲】

【請求項1】

第1の言語により記述された検索条件から第2の言語により記述された電子文書を検索する文書検索装置において、

前記第1の言語により記述された検索条件に合致する電子文書を前記第1の言語により記述された電子文書群の中から検索する第1の検索手段と、

前記第1の検索手段により検索された電子文書の対訳である電子文書を前記第2の言語により記述された電子文書群の中から検索する第2の検索手段と、

前記第2の検索条件により検索されたそれぞれの電子文書に共通して含まれるキーワードであって、前記第2の検索条件により検索された電子文書のみに含まれ、前記第2の言語により記述された電子文書群におけるその他の電子文書には含まれないキーワードを抽出し、抽出されたキーワードにもとづいて前記第2の言語により記述された検索条件を生成する生成手段と、

前記生成手段により生成された検索条件に合致する電子文書を前記第2の言語により記述された電子文書の中から検索する第3の検索手段と、

を備えたことを特徴とする文書検索装置。

【請求項2】

前記第1の検索手段は、前記電子文書のうちその本文が前記検索条件を構成する自然文と意味的に類似する電子文書を前記検索条件に合致する電子文書として検索することを特徴とする前記請求項1に記載の文書検索装置。

【請求項 3】

前記第 2 の検索手段は、前記第 1 の検索手段により検索された電子文書のすべてについて、その対訳である電子文書を前記第 2 の言語により記述された電子文書の中から検索することを特徴とする前記請求項 1 または請求項 2 に記載の文書検索装置。

【請求項 4】

前記第 2 の検索手段は、前記第 1 の検索手段により検索された電子文書のうち一部の電子文書であって、前記第 1 の検索手段による結果に基づいて、前記検索条件に対する所定の合致度以上の電子文書について、その対訳である電子文書を前記第 2 の言語により記述された電子文書の中から検索することを特徴とする前記請求項 1 または請求項 2 に記載の文書検索装置。

10

【請求項 5】

さらに、前記第 1 の検索手段により検索された電子文書を特定できる情報および前記第 3 の検索手段により検索された電子文書を特定できる情報を表示する表示手段を備えたことを特徴とする前記請求項 1 ~ 請求項 4 のいずれか一つに記載の文書検索装置。

【請求項 6】

第 1 の言語により記述された検索条件から第 2 の言語により記述された電子文書をコンピュータを用いて検索する文書検索方法において、

前記コンピュータが、

前記第 1 の言語により記述された検索条件に合致する電子文書を、前記第 1 の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第 1 の検索工程と、

20

前記第 1 の検索工程で検索された電子文書の対訳である電子文書を、前記第 2 の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第 2 の検索工程と、

前記第 2 の検索条件により検索されたそれぞれの電子文書に共通して含まれるキーワードであって、前記第 2 の検索条件により検索された電子文書のみに含まれ、前記第 2 の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群におけるその他の電子文書には含まれないキーワードを抽出し、抽出されたキーワードにもとづいて前記第 2 の言語により記述された検索条件を生成する生成工程と、

前記生成工程で生成された検索条件に合致する電子文書を、前記第 2 の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第 3 の検索工程と、

30

を実行することを特徴とする文書検索方法。

【請求項 7】

前記請求項 6 に記載された方法を前記コンピュータに実行させるプログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

この発明は、ある言語により記述された検索条件からそれとは別の言語により記述された電子文書を検索する文書検索装置、文書検索方法およびその方法をコンピュータに実行させるプログラムに関する。

40

【0002】

【従来の技術】

一般に「多言語文書検索」などと呼ばれる、検索対象文書の言語と検索条件の言語とが異なる文書検索、たとえば英語で記述された文書を日本語のキーワードから検索することは従来から可能であった。

【0003】

一例として、サイバースペース研究所は「TITAN」、AltaVista Company は「AltaVista」という名称で、各国語によるWEB文書の検索サービスをすでに実用化している。また、研究論文としては「AMFにおける多言語によるインター

50

ネット情報検索共同研究プロジェクト」(NTT NEWS RELEASE 1999/02/24)などがある。

【0004】

【発明が解決しようとする課題】

しかしながら、上記従来技術ではいずれも検索条件を検索対象文書の言語に変換するか、あるいは逆に検索対象文書を検索条件の言語に変換するかして、いったん両者の言語を共通化した上で検索をおこなっていた。

【0005】

そして、この変換のためには複数言語間の翻訳システムや、少なくとも単語レベルでの言語置換システムなどが必要であり、翻訳辞書などのデータの準備・洗練コストが大きいほか、検索実行時のシステムにかかる負荷も大きくなってしまふ。しかも、多義的な語は翻訳の過程で意味や概念のズレを生ずることが多いため、検索結果に操作者の意図しないノイズが混入しやすく、処理の複雑さ・煩雑さに見合うだけの検索精度が得られないという問題点があった。

【0006】

なお、大量の文書を統計的に処理することで、辞書を使用せずに複数言語の単語間の対応を特定する試みもなされているが(特開2001-43236)、複雑で大がかりな処理が必要なうえ、現在の技術レベルでは人手で作成された辞書ほどの正確さは期待できない。

【0007】

この発明は上記従来技術による問題点に鑑みてなされたものであって、任意の言語からの任意の言語の文書の検索を簡易な処理で、かつ精度よくおこなうことが可能な文書検索装置、文書検索方法およびその方法をコンピュータに実行させるプログラムを提供することを目的とする。

【0008】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項1に記載の発明にかかる文書検索装置は、第1の言語により記述された検索条件から第2の言語により記述された電子文書を検索する文書検索装置において、前記第1の言語により記述された検索条件に合致する電子文書を前記第1の言語により記述された電子文書の中から検索する第1の検索手段と、前記第1の検索手段により検索された電子文書の対訳である電子文書を前記第2の言語により記述された電子文書の中から検索する第2の検索手段と、前記第2の検索条件により検索されたそれぞれの電子文書に共通して含まれるキーワードを抽出し、抽出されたキーワードにもとづいて前記第2の言語により記述された検索条件を生成する生成手段と、前記生成手段により生成された検索条件に合致する電子文書を前記第2の言語により記述された電子文書の中から検索する第3の検索手段と、を備えたことを特徴とする。

【0009】

この請求項1に記載の発明によれば、多言語文書検索を実現しながらも、検索条件 - 検索対象文書間の言語の差異を解消するための複雑な作業(たとえば機械翻訳など)は一切おこなわれなない。

【0010】

また、請求項2に記載の発明にかかる文書検索装置は、前記請求項1に記載の発明において、前記第1の検索手段が、前記電子文書のうちその本文が前記検索条件を構成する自然文と意味的に類似する電子文書を前記検索条件に合致する電子文書として検索することを特徴とする。

【0011】

この請求項2に記載の発明によれば、分野や話題など、本文の全体としての大意が検索条件と合致する文書のみが検索される。

【0012】

また、請求項3に記載の発明にかかる文書検索装置は、前記請求項1または請求項2に記

10

20

30

40

50

載の発明において、前記第 2 の検索手段が、前記第 1 の検索手段により検索された電子文書のすべてについて、その対訳である電子文書を前記第 2 の言語により記述された電子文書の中から検索することを特徴とする。

【 0 0 1 3 】

この請求項 3 に記載の発明によれば、第 2 の言語による検索条件は、第 1 の言語による検索で得られた文書の対訳が漏れなく使用されて生成される。

【 0 0 1 4 】

また、請求項 4 に記載の発明にかかる文書検索装置は、前記請求項 1 または請求項 2 に記載の発明において、前記第 2 の検索手段が、前記第 1 の検索手段により検索された電子文書のうち一部の電子文書であって、前記第 1 の検索手段による結果に基づいて、前記検索条件に対する所定の合致度以上の電子文書について、その対訳である電子文書を前記第 2 の言語により記述された電子文書の中から検索することを特徴とする。

10

【 0 0 1 5 】

この請求項 4 に記載の発明によれば、第 2 の言語による検索条件は、第 1 の言語による検索で得られた文書のうち、たとえば検索条件との合致度のとくに高かったものの対訳のみが選択的に使用されて生成される。

【 0 0 1 6 】

また、請求項 5 に記載の発明にかかる文書検索装置は、前記請求項 1 ~ 請求項 4 のいずれか一つに記載の発明において、さらに、前記第 1 の検索手段により検索された電子文書を特定できる情報および前記第 3 の検索手段により検索された電子文書を特定できる情報を表示する表示手段を備えたことを特徴とする。

20

【 0 0 1 7 】

この請求項 5 に記載の発明によれば、第 1 の言語による検索の結果と第 2 の言語による検索の結果とがあわせて画面表示される。

【 0 0 1 8 】

また、請求項 6 に記載の発明にかかる文書検索方法は、第 1 の言語により記述された検索条件から第 2 の言語により記述された電子文書をコンピュータを用いて検索する文書検索方法において、前記コンピュータが、前記第 1 の言語により記述された検索条件に合致する電子文書を、前記第 1 の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第 1 の検索工程と、前記第 1 の検索工程で検索された電子文書の対訳である電子文書を、前記第 2 の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第 2 の検索工程と、前記第 2 の検索条件により検索されたそれぞれの電子文書に共通して含まれるキーワードを抽出し、抽出されたキーワードにもとづいて前記第 2 の言語により記述された検索条件を生成する生成工程と、前記生成工程で生成された検索条件に合致する電子文書を、前記第 2 の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第 3 の検索工程と、を実行することを特徴とする。

30

【 0 0 1 9 】

この請求項 6 に記載の発明によれば、多言語文書検索を実現しながらも、検索条件 - 検索対象文書間の言語の差異を解消するための複雑な作業（たとえば機械翻訳など）は一切おこなわれない。

40

【 0 0 2 0 】

また、請求項 7 に記載の発明にかかるプログラムは、前記請求項 6 に記載された方法を前記コンピュータに実行させることを特徴とする。

【 0 0 2 1 】

この請求項 7 に記載の発明によれば、前記請求項 6 に記載された方法がコンピュータにより実行される。

【 0 0 2 2 】

【発明の実施の形態】

以下に添付図面を参照して、この発明による文書検索装置、文書検索方法およびその方法

50

をコンピュータに実行させるプログラムの好適な実施の形態を詳細に説明する。

【0023】

(発明の基本原理)

具体的な実施の形態の説明に入る前に、まず本発明の基本原理について簡単に説明する。図1は、本発明の基本原理を模式的に示す説明図である。本発明における検索対象文書群は、たとえばインターネットから収集された多数のWEB文書であるものとする。図中、Jで始まるIDを付された文書は日本語で記述された文書、Eで始まるIDを付された文書は英語で記述された文書を、それぞれ示している。

【0024】

ここで、グローバルな規模で活動する企業やe-マーケットプレイスに出店する企業などのホームページは、日本語版や英語版など複数の言語のバージョンで作成されていることがある。図中、文書J-1はこうしたホームページの日本語版、文書E-1は同じページの英語版を、それぞれ示している。文書J-1と文書E-1とは、記述内容は同一でただ当該内容を記述する言語が異なるのみである。

【0025】

この文書J-1のように忠実な英語訳を有する日本語文書、あるいは文書E-1のように忠実な日本語訳を有する英語文書を、以下では「対訳つき文書(群)」と総称する。これに対し、日本語版しかない文書J-2やJ-3、あるいは逆に英語版しかない文書E-2、E-3、E-4などを、以下では「対訳なし文書(群)」と総称する。

【0026】

検索対象文書のすべてが対訳つき文書であれば、容易に任意の検索条件による他言語の文書の検索を実現することができる。すなわち、たとえば日本語の検索条件から英語の文書を検索できるようにするためには、当該日本語の検索条件に合致した日本語文書でなく、当該日本語文書に対応する英語文書を検索結果として返すようにすればよい。

【0027】

しかしながら、実際の検索対象文書には図示するように対訳つき文書と対訳なし文書とが混在しており、少なくとも後者については、上述のように検索条件側の言語を検索対象文書側の言語に合わせるか、逆に検索対象文書側の言語を検索条件側の言語に合わせるかした上で検索をおこなうのが従来の手法であった。

【0028】

これに対し、本発明では以下で詳述するように、日本語の検索条件に合致した日本語文書に対応する英語文書の本文を英語の検索条件とみなし、当該条件に合致する英語文書を検索結果として返すことで、複雑で困難な言語の変換処理や統計処理を介することなく、日本語の検索条件からの英語文書の検索をおこなう。

【0029】

すなわち、図1に模式的に示すように、まず操作者から日本語で入力された自然文を検索条件として、当該条件に合致する(当該自然文に全体として類似する、と言ってもよい)日本語文書を検索する(図中 1)。

【0030】

そして、日本語文書J-1とJ-2とが上記条件に合致したものとすると、これらの適合文書のうち日本語文書J-1には対応する英語文書E-1が存在するので、つぎにこの英語文書E-1の本文である英語の自然文を新たな検索条件として、当該条件に合致する(当該自然文に全体として類似する、と言ってもよい)英語文書を検索する(図中 2)。

【0031】

そして、上記新たな条件に合致する文書、すなわち文書E-1そのものと、文書E-1に類似する文書E-2およびE-3を、最終的な検索結果として操作者に提示する。

【0032】

すなわち本発明では、当初の検索条件の日本語を自前で英語に翻訳するのではなく、当該日本語に類似する日本語文書(この例では文書J-1)につきすでに人手で作成されてい

10

20

30

40

50

る、正確な英語訳（文書 E - 1）を上記条件の英語訳に相当すると便宜上みなして、これを新たな検索条件として英語文書の検索をおこなうわけである。

【 0 0 3 3 】

日本語の検索条件に合致した日本語文書の忠実な英語訳は、当該検索条件となった自然文と内容的にも言語的にも同一ではないものの、内容的に類似はしている。少なくとも、従来の機械翻訳技術で自動生成される英語訳よりは、日本語で記述された当初の検索条件からの意味的・概念的なズレが小さく、これを検索条件として検索をおこなうことにより、最終的な検索結果中に含まれるノイズを減少させることができる。

【 0 0 3 4 】

また、人手による対訳は自動生成された対訳よりも自然な（すなわち、ネイティブスピーカーが読んでも違和感のない質を備えた）文章であり、語用法や論理の展開方法もその言語に即したものが使用されるので、これを検索条件として採用することで、同じ言語で記述された検索対象文書との類似性をより正確に判定できると考えられる。もっとも、踏み台となる対訳は必ずしも人手により作成されたものでなくともよく、原文とのズレが大きくなければ機械翻訳されたものであっても構わない（人手で作成された対訳であれば通常上記のようなメリットもある、というだけのことである）。

【 0 0 3 5 】

このように本発明では、 1 操作者から入力された、日本語の自然文を検索条件とする日本語文書の検索、 2 当該検索で得られた日本語文書に対応する英語文書の本文である、英語の自然文を検索条件とする英語文書の検索、を連鎖的におこなうことで、結果的に日本語で記述された検索条件から、英語により記述された文書を検索することが可能となる。

【 0 0 3 6 】

（発明の実施の形態）

つぎに、図 2 は本発明の実施の形態による文書検索装置のハードウェア構成を示す説明図である。同図において、201 は装置全体を制御する CPU を、202 は基本入出力プログラムを記憶した ROM を、203 は CPU 201 のワークエリアとして使用される RAM を、それぞれ示している。

【 0 0 3 7 】

また、204 は CPU 201 の制御にしたがって HD（ハードディスク）205 に対するデータのリード/ライトを制御する HDD（ハードディスクドライブ）を、205 は HDD 204 の制御にしたがって書き込まれたデータを記憶する HD を、それぞれ示している。

【 0 0 3 8 】

また、206 は CPU 201 の制御にしたがって FD（フロッピーディスク）207 に対するデータのリード/ライトを制御する FDD（フロッピーディスクドライブ）を、207 は FDD 206 の制御にしたがって書き込まれたデータを記憶する着脱自在の FD を、それぞれ示している。

【 0 0 3 9 】

また、208 はカーソル、メニュー、ウィンドウ、あるいは文字や画像などの各種データを表示するディスプレイを、209 は通信ケーブル 210 を介して LAN などのネットワークに接続され、当該ネットワークと CPU 201 とのインターフェースとして機能するネットワーク I/F を、それぞれ示している。

【 0 0 4 0 】

また、211 は文字、数値、各種指示などの入力のための複数のキーを備えたキーボードを、212 は各種指示の選択や実行、処理対象の選択、カーソルの移動などをおこなうマウスを、それぞれ示している。また、213 は着脱可能な記録媒体である CD-ROM を、214 は CD-ROM 213 に対するデータのリードを制御する CD-ROM ドライブを、200 は上記各部を接続するためのバスまたはケーブルを、それぞれ示している。

【 0 0 4 1 】

10

20

30

40

50

つぎに、図3は本発明の実施の形態による文書検索装置の機能的構成を示す説明図である。図示するように、本発明による文書検索装置は文書記憶部300、日本語検索条件入力部301、日本語文書検索部302、英語検索条件生成部303、英語文書検索部304および検索結果表示部305を含む構成である。

【0042】

まず、文書記憶部300は後述する日本語文書検索部302および英語文書検索部304による検索対象となる文書群を保持する機能部である。ここでは、文書記憶部300内の文書はインターネットから収集された多数のWEB文書であるものとし、そのうち日本語で記述されたものは日本語文書記憶部300aに、英語で記述されたものは英語文書記憶部300bに、それぞれ保持されるものとする。

10

【0043】

なお、上述のように文書記憶部300内の文書は、一部が対訳つき文書であり残りは対訳なし文書である。そして、対訳つき文書はその属性情報(付属情報)として、他言語で記述された対訳文書のID(IDに限らず、当該文書を特定できる情報であれば何でもよい)を保持している。

【0044】

たとえば、日本語文書J-1と英語文書E-1とが対訳関係にあれば、前者の属性情報には後者のIDである「E-1」が、後者の属性情報には前者のIDである「J-1」が、それぞれあらかじめ書き込まれている。

20

【0045】

つぎに、図4は本発明の実施の形態による文書検索装置の、文書検索処理の手順を示すフローチャートである。以下、同図に示す手順に沿って、図3に示した残りの各部の機能を順次説明する。

【0046】

ステップS401で、本発明による文書検索装置の日本語検索条件入力部301は、図5に示すような検索条件入力画面をディスプレイ208に表示して操作者からの入力待ちとなる。そして、キーボード211などから入力された文字を検索条件入力エリア500内に順次表示する。

【0047】

なお、図示するようにここでは検索条件として複数の文からなる自然文が入力されたものとするが、単一の文からなる自然文、単数あるいは複数のキーワードなど、日本語の文字列であればどのようなものであってもよい。

30

【0048】

つぎにステップS402で、検索を実行すべき旨の指示が入力されたこと、すなわち図5に示す検索実行ボタン501がマウス212でクリックされたことを検知すると(ステップS402:Yes)、日本語検索条件入力部301はその時点での上記入力エリア500内の文字列を検索条件として、後述する日本語文書検索部302に引き渡す。

【0049】

そして、これを受けた日本語文書検索部302は、ステップS403で上記検索条件により日本語文書記憶部300aを検索する。この日本語文書検索部302による検索手法は、ある言語で記述された検索条件から当該言語により記述された文書を検索できるもの(単一言語内での文書検索が可能なもの、と言ってもよい)であれば何であっててもよいが、ここでは一般に「ベクトル空間法」と呼ばれる手法を採用する。

40

【0050】

「ベクトル空間法」とは、検索条件の特徴ベクトルと、検索対象となる個々の文書の特徴ベクトルとのコサイン距離をそれぞれ計算し、この距離が絶対的または相対的に小さい文書を、検索条件に合致する適合文書として操作者に提示するものである。

【0051】

ここでの特徴ベクトルとは、n個のキーワード(語彙)に対応するn個の要素値からなるn次元のベクトルであって、個々の要素値は最も単純には、対応するキーワードの出現頻

50

度により決定される。たとえば、本文内に一つのキーワードしか含まない文書の特徴ベクトルは、(0、1、0、0、・・・)のように当該キーワードに対応する要素の値だけが1で、残りn-1個の要素値がすべて0となるようなベクトルである。

【0052】

このベクトル空間法では、本文内に出現するキーワードの全体としての傾向が検索条件と類似するような文書ほど検索条件との距離が小さくなり、したがって適合文書とされる可能性が高くなる。そのため、検索条件中の特定のキーワードが含まれるか否かにより単純に文書を選別するブーリアン検索(一般のキーワード検索)に比べ、検索結果中のノイズが少ないという利点がある。

【0053】

日本語文書検索部302は、上記距離を基礎として検索対象文書の順位づけ、あるいは得点づけをおこない、最高順位/最高得点の文書から一定数の文書、あるいは所定の順位/所定の得点以上のすべての文書など、検索条件に対する合致度の高い文書を適合文書とする。そして、これら適合文書のID(IDに限らず、当該文書を特定できる情報であれば何であってもよい)を、後述する英語検索条件生成部303および検索結果表示部305にそれぞれ引き渡す。

【0054】

図4の手順に戻り、つぎにステップS404で、日本語文書検索部302からその検索結果を引き渡された英語検索条件生成部303は、引き渡されたIDで特定される各文書の属性情報を日本語文書記憶部300aから読み出す。そして、その中に英語文書のIDが一つでも含まれているかどうか、すなわち上記検索で拾い出された日本語文書の中に、一つでも対訳つき文書が含まれているかどうかを判定する。

【0055】

そして、上記結果中に一つでも対訳つき文書が含まれていれば(ステップS404:Yes)、つぎにステップS405で、それぞれの対訳つき文書に対応する英語文書の本文、すなわち上記で読み出した属性情報中の各IDにより特定される英語文書の本文を、英語文書記憶部300bから順次読み出す。そして、これらの英語の自然文から、後述する英語文書検索部304に与えるための検索条件を生成する。

【0056】

なお、ここでは英語検索条件生成部303は、日本語文書検索部302による検索結果中のすべての対訳つき文書について当該対訳を読み出すようにしたが、一部の対訳つき文書を選択してその対訳のみを読み出すようにしてもよい。

【0057】

たとえば、適合文書のうち最高順位/最高得点の文書から一定数の文書、あるいは所定の順位/所定の得点以上のすべての文書など、検索条件に対する合致度のとくに高い文書に限って、その対訳を英語の検索条件として採用する。逆に言えば、適合文書であっても検索条件に対する合致度が低い文書については、対訳が存在していてもその存在を無視する。

【0058】

このように、適合文書の中でもとくにレベルの高い文書の対訳を採用することで、日本語検索条件入力部301から入力された日本語の検索条件と、英語検索条件生成部303で生成される英語の検索条件とのズレが少なくなり、最終的な検索結果はより絞り込まれた、適合率の高いものとなる。もっとも、その反面で再現率は低くなってしまいうので、漏れない検索が必要であればここでの例のように、日本語の適合文書について存在するすべての対訳を英語の検索条件として採用すればよい。

【0059】

なお、検索条件として採用された英語文書が複数ある場合、英語検索条件生成部303は各文書の本文を結合して一続きの自然文とした上で、後述する英語文書検索部304に引き渡す。そして、これを受けた英語文書検索部304では、この自然文に全体として類似する英語文書を検索することになる。

10

20

30

40

50

【 0 0 6 0 】

もっとも、採用された文書ごとにその本文を一つの検索条件とみなして、それぞれ別個に英語文書検索部 3 0 4 に引き渡し、上記文書の個数分だけ同様の検索を繰り返させるようにしてもよい。この場合、後述する検索結果表示部 3 0 5 では、各条件により検索された英語文書を区別して表示したり、あるいは各条件により検索された英語文書の和集合を取った上でまとめて表示したりすることが可能である。

【 0 0 6 1 】

なお、上記で採用されたそれぞれの文書（あるいは少なくともその多くの文書）に共通して含まれるキーワードや、採用された文書のみに含まれ、それ以外の文書には含まれないようなキーワード（採用された文書群をその母体となった文書群全体に対して特徴づけるようなキーワード）のみを特定して、これらのキーワードから検索条件を生成するようにしてもよい。

10

【 0 0 6 2 】

つぎに、英語検索条件生成部 3 0 3 から上記検索条件を引き渡された英語文書検索部 3 0 4 は、ステップ S 4 0 6 で英語文書記憶部 3 0 0 b を検索し、上記条件に合致した英語文書の ID（ID に限らず、当該文書を特定できる情報であれば何であってもよい）を、後述する検索結果表示部 3 0 5 に引き渡す。

【 0 0 6 3 】

英語文書検索部 3 0 4 による英語文書の検索は、日本語文書検索部 3 0 2 による日本語文書の検索と同様、ここではベクトル空間法によるものとする。ただし、必ずしもこの手法に限定されるものではなく、また両機能部による検索が本質的に同一である必要もない。たとえば、日本語文書検索部 3 0 2 は上述のベクトル空間法により、検索条件である自然文に概ね類似する文書の検索をおこない、英語文書検索部 3 0 4 はブーリアン検索により、検索条件として採用された英語文書内のキーワードを確実に含む文書のみを検索をおこなうようにしてもよい。

20

【 0 0 6 4 】

つぎに、日本語文書検索部 3 0 2 から検索結果の日本語文書の各 ID、英語文書検索部 3 0 4 から検索結果の英語文書の各 ID をそれぞれ引き渡された検索結果表示部 3 0 5 は、ステップ S 4 0 7 で図 6 に示すような検索結果表示画面をディスプレイ 2 0 8 に表示する。

30

【 0 0 6 5 】

同図において、日本語文書表示フレーム 6 0 0 には日本語文書検索部 3 0 2 により検索された日本語文書の各見出し、英語文書表示フレーム 6 0 1 には英語文書検索部 3 0 4 により検索された英語文書の各見出しが、それぞれ検索条件に対する合致度の高い順に表示される。この見出しをマウス 2 1 2 でクリックすると、当該見出しを有する文書の本文を表示させることができる。なお、同図では見出しの横の括弧内に文書の ID をあわせて表示しているが、これはあってもなくてもよい。

【 0 0 6 6 】

以上説明した実施の形態によれば、ある言語により記述された文書を、それとは別の言語による検索条件で検索することが可能でありながら、その過程において複雑な翻訳処理や統計処理などは一切発生せず、従来技術に比較してシステムにかかる負荷が格段に小さい。

40

【 0 0 6 7 】

また、従来多大な時間と労力とを要していた、言語間の翻訳のための辞書類の整備が不要であり、ただ一部に対訳つき文書を含む文書群が収集できさえすればよい。そして、近年では WEB 文書を始めとして、あらかじめ対訳つきで作成されている電子文書が少なくないので、この収集も容易である。逆に言えば、本発明は収集した文書群にしばしば対訳つき文書が含まれることに注目して、この状況を利用し、これを足がかりとして多言語文書検索が実現できないかとの着想を得たものである。

【 0 0 6 8 】

50

なお、上述した実施の形態ではインターネット上のWEB文書の検索を例としたが、このほか国際的企業の社内ネットワークにおけるFAQ文書の検索、各種研究・教育機関における各国語による学术论文の検索などにも本発明は応用可能である。

【0069】

また、上述した実施の形態では日本語から英語の文書を検索するようにしたが、逆に英語から日本語の文書を検索することも可能なことは言うまでもない。なお、図7に模式的に示すように、たとえば日本語で検索された日本語文書から対訳関係にある英語文書を取得し、当該英語文書の本文で検索された英語文書から対訳関係にあるドイツ語文書を取得し、さらに当該ドイツ語文書の本文により検索されたドイツ語文書を検索結果とすれば、結果的に日本語の検索条件からドイツ語の文書が検索されたことになり、このように対訳関係にある文書を複数言語にわたって芋づる式にたどってゆくことで、あらゆる言語からのあらゆる言語の文書の検索が可能となる。

10

【0070】

なお、図1や図7では対訳文書と当該対訳文書により検索される文書群とが、常に同一のデータベース内に存在するかのよう描いているが、必ずしも対訳文書の抽出の母体となった文書群に対して当該対訳文書による検索をおこなわなければならないものではない。すなわち、たとえば日本語の検索条件で検索された日本語文書の対訳をデータベースAから取得し、その本文を英語の検索条件として、それとは別のデータベースBを検索するのであってもよい。

【0071】

20

なお、上述した日本語文書検索部302が請求項にいう「第1の検索手段」に、そのおこなう処理が請求項にいう「第1の検索工程」に、それぞれ相当する。また、英語検索条件生成部303が請求項にいう「第2の検索手段」および「生成手段」を兼ね、そのおこなう処理に請求項にいう「第2の検索工程」および「生成工程」が含まれる。また、英語文書検索部304が請求項にいう「第3の検索手段」に、そのおこなう処理が請求項にいう「第3の検索工程」に、それぞれ相当する。さらに、検索結果表示部305が請求項にいう「表示手段」に相当する。

【0072】

なお、上述した日本語検索条件入力部301～検索結果表示部305は、それぞれHD205などからRAM203に読み出されたプログラムの命令にしたがってCPU201が命令処理を実行することにより、各部の機能を実現するものである。また、とくに日本語文書検索部302と英語文書検索部304とは、具体的には本出願人が製造・販売する文書検索エンジンの「Concept Base Search」により実現される。

30

【0073】

なお、上記プログラムはHD205のほか、FD207、CD-ROM213あるいはMOなどの各種記録媒体に格納することができ、この媒体により配布することができるほか、ネットワークを介して配布することも可能である。

【0074】

【発明の効果】

以上説明したように請求項1に記載の発明は、第1の言語により記述された検索条件から第2の言語により記述された電子文書を検索する文書検索装置において、前記第1の言語により記述された検索条件に合致する電子文書を前記第1の言語により記述された電子文書の中から検索する第1の検索手段と、前記第1の検索手段により検索された電子文書の対訳である電子文書を前記第2の言語により記述された電子文書の中から検索する第2の検索手段と、前記第2の検索条件により検索されたそれぞれの電子文書に共通して含まれるキーワードを抽出し、抽出されたキーワードにもとづいて前記第2の言語により記述された検索条件を生成する生成手段と、前記生成手段により生成された検索条件に合致する電子文書を前記第2の言語により記述された電子文書の中から検索する第3の検索手段と、を備えたので、多言語文書検索を実現しながらも、検索条件-検索対象文書間の言語の差異を解消するための複雑な作業は一切おこなわれず、これによって、任意の言語から

40

50

の任意の言語の文書の検索を簡易な処理でおこなうことが可能な文書検索装置が得られるという効果を奏する。

【0075】

また、請求項2に記載の発明は、前記請求項1に記載の発明において、前記第1の検索手段が、前記電子文書のうちその本文が前記検索条件を構成する自然文と意味的に類似する電子文書を前記検索条件に合致する電子文書として検索するので、分野や話題など、本文の全体としての大意が検索条件と合致する文書のみが検索され、これによって、任意の言語からの任意の言語の文書の検索を簡易な処理で、かつ精度よくおこなうことが可能な文書検索装置が得られるという効果を奏する。

【0076】

また、請求項3に記載の発明は、前記請求項1または請求項2に記載の発明において、前記第2の検索手段が、前記第1の検索手段により検索された電子文書のすべてについて、その対訳である電子文書を前記第2の言語により記述された電子文書の中から検索するので、第2の言語による検索条件は、第1の言語による検索で得られた文書の対訳が漏れなく使用されて生成され、これによって、任意の言語からの任意の言語の文書の検索を簡易な処理で、かつ精度よく（具体的には、再現率が高い）おこなうことが可能な文書検索装置が得られるという効果を奏する。

【0077】

また、請求項4に記載の発明は、前記請求項1または請求項2に記載の発明において、前記第2の検索手段が、前記第1の検索手段により検索された電子文書のうち一部の電子文書であって、前記第1の検索手段による結果に基づいて、前記検索条件に対する所定の合致度以上の電子文書について、その対訳である電子文書を前記第2の言語により記述された電子文書の中から検索するので、第2の言語による検索条件は、第1の言語による検索で得られた文書のうち、たとえば検索条件との合致度のとくに高かったものの対訳のみが選択的に使用されて生成され、これによって、任意の言語からの任意の言語の文書の検索を簡易な処理で、かつ精度よく（具体的には、適合率が高い）おこなうことが可能な文書検索装置が得られるという効果を奏する。

【0078】

また、請求項5に記載の発明は、前記請求項1～請求項4のいずれか一つに記載の発明において、さらに、前記第1の検索手段により検索された電子文書を特定できる情報および前記第3の検索手段により検索された電子文書を特定できる情報を表示する表示手段を備えたので、第1の言語による検索の結果と第2の言語による検索の結果とが合わせて画面表示され、これによって、多言語文書検索の結果をその中間結果も含めて、分かりやすく操作者に提示することが可能な文書検索装置が得られるという効果を奏する。

【0079】

また、請求項6に記載の発明は、第1の言語により記述された検索条件から第2の言語により記述された電子文書をコンピュータを用いて検索する文書検索方法において、前記コンピュータが、前記第1の言語により記述された検索条件に合致する電子文書を、前記第1の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第1の検索工程と、前記第1の検索工程で検索された電子文書の対訳である電子文書を、前記第2の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第2の検索工程と、前記第2の検索条件により検索されたそれぞれの電子文書に共通して含まれるキーワードを抽出し、抽出されたキーワードにもとづいて前記第2の言語により記述された検索条件を生成する生成工程と、前記生成工程で生成された検索条件に合致する電子文書を、前記第2の言語により記述され、あらかじめ所定の記憶領域に記憶された電子文書群の中から検索する第3の検索工程と、を実行するので、多言語文書検索を実現しながらも、検索条件 - 検索対象文書間の言語の差異を解消するための複雑な作業は一切おこなわれず、これによって、任意の言語からの任意の言語の文書の検索を簡易な処理でおこなうことが可能な文書検索方法が得られるという効果を奏する。

。

10

20

30

40

50

【 0 0 8 0 】

また、請求項 7 に記載の発明によれば、前記請求項 6 に記載された方法を前記コンピュータに実行させることが可能なプログラムが得られるという効果を奏する。

【図面の簡単な説明】

【図 1】本発明の基本原理を模式的に示す説明図である。

【図 2】本発明の実施の形態による文書検索装置のハードウェア構成を示す説明図である。

【図 3】本発明の実施の形態による文書検索装置の機能的構成を示す説明図である。

【図 4】本発明の実施の形態による文書検索装置の文書検索処理の手順を示すフローチャートである。

10

【図 5】本発明の実施の形態による文書検索装置における、検索条件入力画面の一例を示す説明図である。

【図 6】本発明の実施の形態による文書検索装置における、検索結果表示画面の一例を示す説明図である。

【図 7】本発明の他の実施例の基本原理を模式的に示す説明図である。

【符号の説明】

2 0 0 バスまたはケーブル

2 0 1 C P U

2 0 2 R O M

2 0 3 R A M

20

2 0 4 H D D

2 0 5 H D

2 0 6 F D D

2 0 7 F D

2 0 8 ディスプレイ

2 0 9 ネットワーク I / F

2 1 0 通信ケーブル

2 1 1 キーボード

2 1 2 マウス

2 1 3 C D - R O M

30

2 1 4 C D - R O M ドライブ

3 0 0 文書記憶部

3 0 0 a 日本語文書記憶部

3 0 0 b 英語文書記憶部

3 0 1 日本語検索条件入力部

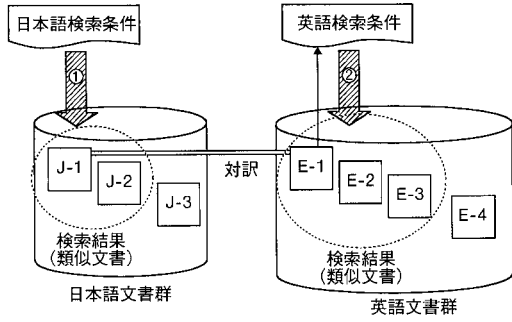
3 0 2 日本語文書検索部

3 0 3 英語検索条件生成部

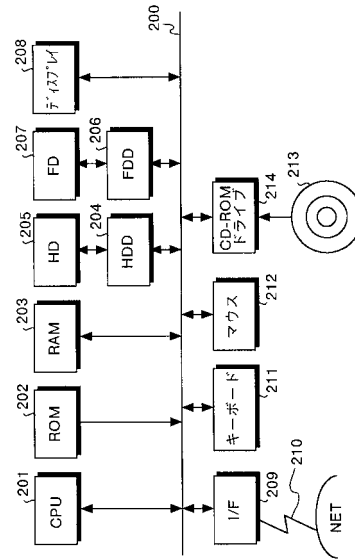
3 0 4 英語文書検索部

3 0 5 検索結果表示部

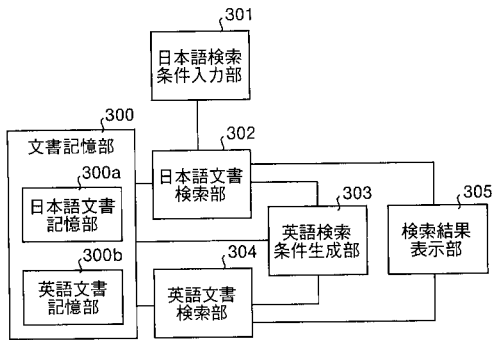
【図1】



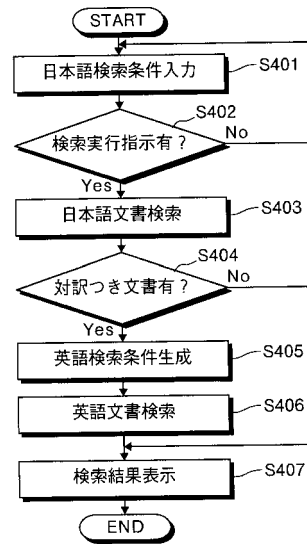
【図2】



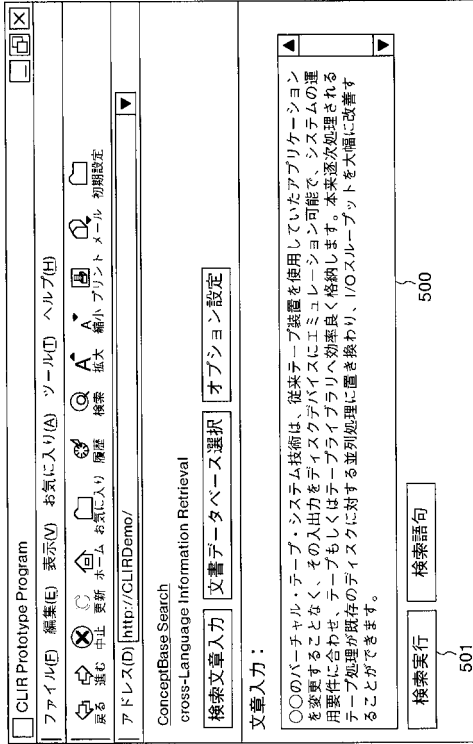
【図3】



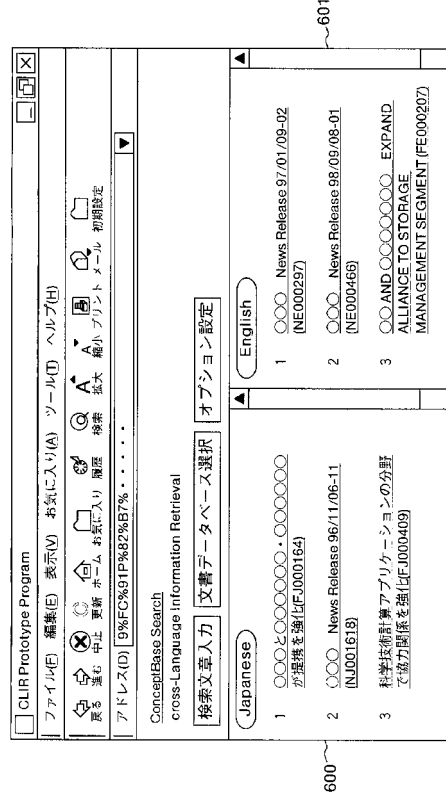
【図4】



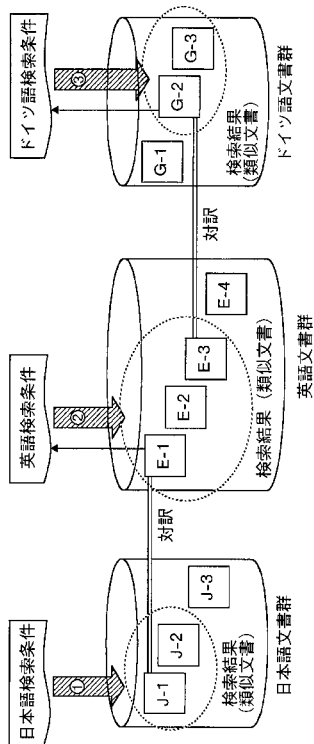
【 図 5 】



【 図 6 】



【 図 7 】



フロントページの続き

合議体

審判長 田口 英雄

審判官 和田 財太

審判官 小曳 満昭

(56)参考文献 特開2000-20524(JP,A)
特開平10-289244(JP,A)

(58)調査した分野(Int.Cl., DB名)
G06F17/27-17/30