(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4717821号
(P4717821)

(45)発行日　平成23年7月6日(2011.7.6)　(24)登録日　平成23年4月8日(2011.4.8)

(51)Int.Cl.　　F I
G06F 17/30 (2006.01)　　G06F 17/30 320C
G06F 17/28 (2006.01)　　G06F 17/30 170A
　　G06F 17/28 Z

請求項の数 15　(全 24 頁)

(21)出願番号　特願2006-533909 (P2006-533909)
(86)(22)出願日　平成16年9月13日 (2004.9.13)
(65)公表番号　特表2007-507796 (P2007-507796A)
(43)公表日　平成19年3月29日 (2007.3.29)
(86)国際出願番号　PCT/US2004/029772
(87)国際公開番号　WO2005/033967
(87)国際公開日　平成17年4月14日 (2005.4.14)
審査請求日　平成19年8月20日 (2007.8.20)
(31)優先権主張番号　10/676,724
(32)優先日　平成15年9月30日 (2003.9.30)
(33)優先権主張国　米国(US)

(73)特許権者　502208397
グーグル　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９４０４３　マウンテン　ビュー　アンフィシアター　パークウェイ　１６００
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100062409
弁理士　安村　高明
(74)代理人　100113413
弁理士　森下　夏樹
(72)発明者　ミッタル，　ビブ
アメリカ合衆国　カリフォルニア　９４０８７，　サニーベール，　エルソナ　ドライブ　１３２７

最終頁に続く

(54)【発明の名称】ターゲットページとは異なる文字セットおよび／または言語で書かれたクエリを使用する検索のための方法

(57)【特許請求の範囲】
【請求項１】
複数のサーバと１つの入力デバイスとを含むシステムにおける方法であって、該複数のサーバの各々はプロセッサを含み、該方法は、
該複数のサーバのうちの１つのサーバが、第１のフォーマットで書かれ、所定の言葉を含む第１のアンカーテキストのセットを識別することと、
該複数のサーバのうちの該１つのサーバが、該第１のアンカーテキストのセットが指す文書のセットを識別することと、
該複数のサーバのうちの該１つのサーバが、第２のフォーマットで書かれ、該識別された文書のセットを指す第２のアンカーテキストのセットを識別することと、
該複数のサーバのうちの該１つのサーバの該プロセッサが、該第１のフォーマットにおける該所定の言葉の表示が該第２のフォーマットにおける該所定の言葉の表示に対応することを決定するために該第２のアンカーテキストのセットを分析することであって、該第２のアンカーテキストのセットを分析することは、該所定の言葉が該第２のアンカーテキストのセット内の言葉に対応する確率を計算することを含む、ことと
を包含する、方法。
【請求項２】
前記第１のフォーマットは第１の文字セットを含み、前記第２のフォーマットは第２の文字セットを含む、請求項１に記載の方法。
【請求項３】

前記第１のフォーマットは第１の言語を含み、前記第２のフォーマットは第２の言語を含む、請求項１に記載の方法。
【請求項４】
前記第２のアンカーテキストのセットを分析することは、該第２のアンカーテキストのセットで最も頻繁に現われる言葉を識別することと、該第２のフォーマットにおいて、該最も頻繁に現われる言葉を前記所定の言葉の表示として指定することとを含む、請求項１に記載の方法。
【請求項５】
前記確率は、ベイズ法、ヒストグラムスムージング、カーネルスムージング、および縮小推定量のうちの少なくとも一つを用いて得られる、請求項１に記載の方法。
【請求項６】
前記所定の言葉が前記第２のアンカーテキストのセットにおける言葉に対応する前記確率は、該第２のアンカーテキストのセットにおける該言葉の発生回数を、該第２のアンカーテキストのセットにおける全言葉の総発生回数で割ることによって得られる、請求項１に記載の方法。
【請求項７】
前記第２のアンカーテキストのセットを分析することは、前記所定の言葉が該第２のアンカーテキストのセットにおけるそれぞれの言葉に対応する確率を計算することを包含する、請求項１に記載の方法。
【請求項８】
前記第２のアンカーテキストのセットを分析することは、該第２のアンカーテキストのセットにおいて最も頻繁に現われる言葉を識別することを包含する、請求項１に記載の方法。
【請求項９】
前記第１のフォーマットは、ローマ字、ｒｏｍａｊａ、およびピンインから成る群から選択され、前記第２の文字セットは、カタカナ、ひらがな、漢字、ハングル、ハンジャ、および伝統的な中国文字から成る群から選択される、請求項２に記載の方法。
【請求項１０】
前記文書がウェブページを含む、請求項１に記載の方法。
【請求項１１】
前記複数のサーバのうちの前記１つのサーバが、前記第１のフォーマットで書かれ、前記所定の言葉を含むクエリを前記入力デバイスから得ることと、
該複数のサーバのうちの該１つのサーバが、少なくとも部分的に前記分析することに基づいて、該クエリを前記第２のフォーマットに変換することと、
該複数のサーバのうちの該１つのサーバが、データベースで、該変換されたクエリに応じた、該第２のフォーマットで書かれた情報を検索することと
をさらに包含する、請求項１に記載の方法。
【請求項１２】
前記第１のアンカーテキストのセットを識別すること、前記文書のセットを識別すること、前記第２のアンカーテキストのセットを識別すること、前記分析すること、前記得ること、前記変換すること、および前記検索することは、列挙された順序で実行される、請求項１１に記載の方法。
【請求項１３】
コンピュータプログラムを格納したコンピュータ読み取り可能格納媒体であって、該コンピュータプログラムは、プロセッサを含むコンピュータシステムによって実行される場合に該コンピュータシステムに動作を実行させるように操作可能な命令を含み、該動作が、
第１のフォーマットで書かれ、所定の言葉を含む第１のアンカーテキストのセットを識別することと、
該第１のアンカーテキストのセットが指すウェブページのセットを識別することと、

第２のフォーマットで書かれ、該識別されたウェブページのセットを指す第２のアンカーテキストのセットを識別することと、
該プロセッサが、該第１のフォーマットにおける該所定の言葉の表示が該第２のフォーマットにおける該所定の言葉の表示に対応する確率を計算することと
を包含する、コンピュータ読み取り可能格納媒体。
【請求項１４】
前記コンピュータシステムによって実行される場合に該コンピュータシステムに動作を実行させるように操作可能な命令をさらに含み、該動作が、
前記プロセッサが、少なくとも部分的に、前記第２のフォーマットにおける前記所定の言葉の表示を含む検索結果をユーザーが選択することの、該第２のフォーマットにおける該所定の言葉の他の表示を含む検索結果をユーザーが選択することに対する比率に基づいて、前記第１のフォーマットにおける該所定の言葉の表示が該第２のフォーマットにおける該所定の言葉の表示に対応する確率を修正することを包含する、請求項１３に記載のコンピュータ読み取り可能格納媒体。
【請求項１５】
前記確率は、少なくとも部分的に、ベイズ法、ヒストグラムスムージング、カーネルスムージング、および縮小推定量のうちの少なくとも一つを用いて計算される、請求項１３に記載のコンピュータ読み取り可能格納媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
（関連出願の参照）
本出願は、２０００年１２月２６日に出願され、「ＭＥＴＨＯＤＳ　ＡＮＤ　ＡＰＰＡＲＡＴＵＳ　ＦＯＲ　ＰＲＯＶＩＤＩＮＧ　ＳＥＡＲＣＨ　ＲＥＳＵＬＴＳ　ＩＮ　ＲＥＳＰＯＮＳＥ　ＴＯ　ＡＮ　ＡＭＢＩＧＵＯＵＳ　ＳＥＡＲＣＨ　ＱＵＥＲＹ」と題された、米国特許出願シリアル番号第０９／７４８，４３１号の一部継続であり、２０００年７月６日に出願され、「ＤＡＴＡ　ＥＮＴＲＹ　ＡＮＤ　ＳＥＡＲＣＨ　ＦＯＲ　ＨＡＮＤＨＥＬＤ　ＤＥＶＩＣＥＳ」と題された、米国特許仮出願シリアル番号第６０／２１６，５３０号の優先権を、米国特許法第１１９条（ｅ）に基づき主張し、その両方は、それらの全体においてここに援用される。
【０００２】
（発明の分野）
本発明は一般に、情報検索に関する。より詳細には、検索される文章の少なくとも一部の文字セットまたは言語とは異なる文字セットまたは言語において書かれたクエリを使用して検索を実行するためのシステムおよび方法が開示される。
【背景技術】
【０００３】
多くの検索エンジンは、エンドユーザが従来のキーボードなどのようなものを用いて検索クエリを入力するという想定の下で動作し、そこで、英数字の入力は難しいことではない。小さなデバイスがより一般的になってはいるが、しかしながら、この想定はいつも有効とは限らない。例えば、ユーザは、ＷＡＰ（ワイヤレス・アプリケーション・プロトコル）規格をサポートする携帯電話を使用して、検索エンジンにクエリし得る。携帯電話などのデバイスは通常、データ入力インターフェースを有し、ユーザによる特定のアクション（例えばキーを押すなど）が一つ以上の英数字文字に対応し得る。ＷＡＰ構成の詳細は、ｈｔｔｐ：／／ｗｗｗ１．ｗａｐｆｏｒｕｍ．ｏｒｇ／ｔｅｃｈ／ｄｏｃｕｍｅｎｔｓ／ＳＰＥＣ－ＷＡＰＡｒｃｈ－１９９８０４３９．ｐｄｆ（「ＷＡＰ　１００　Ｗｉｒｅｌｅｓｓ　Ａｐｐｌｉｃａｔｉｏｎ　Ｐｒｏｔｏｃｏｌ　Ａｒｃｈｉｔｅｃｔｕｒｅ　Ｓｐｅｃｉｆｉｃａｔｉｏｎ」）にて利用可能である。
【０００４】
通常の場合、ＷＡＰユーザは、検索クエリのページにナビゲートされ、ユーザが検索ク

エリを入力するフォームを提示される。従来の方法では、ユーザは、多数のキーを押して、特定の文字を選択し得る。標準の電話のキーパッドでは、ユーザは、例えば、文字「ｂ」を選択する場合、「２」のキーを２回押す。または、文字「ｓ」を選択したい場合は、「７」のキーを４回押す。従って、「ｂｅｎ　ｓｍｉｔｈ」というクエリを入力するには、ユーザは通常、２２３３６６０７７７７６４４４８４４という一連のキーを押して入力する必要があり、以下のような文字に対応する。

２２→ｂ
３３→ｅ
６６→ｎ
０→スペース
７７７７→ｓ
６→ｍ
４４４→ｉ
８→ｔ
４４→ｈ

ユーザが検索リクエストを入力した後、検索エンジンは、ユーザから文字を受け取り、あたかも、ユーザが従来のキーボードを用いて、デスクトップのブラウザからリクエストを受け取ったかのように、同様の方法で処理する。

【０００５】

前述の例から理解できるように、データ入力のこの形式は、「ｂｅｎ　ｓｍｉｔｈ」に対応する９つの英数字文字（スペースを含む）を入力するために１８回もキー入力を必要とする点で、非効率的である。

【０００６】

同様の困難さは、ターゲットでない（ｎｏｎ－ｔａｒｇｅｔ）言語のキーボードを用いてクエリをタイピングする場合に生じ得る。例えば、日本語のテキストは、ひらがな、カタナカ、および漢字などを含む様々な異なる文字セットを用いて表現され得、そのどれもが、ローマ字（Ｒｏｍａｎ　ａｌｐｈａｂｅｔ）に基づいた通常のＡＳＣＩＩキーボードを用いて容易に入力されるものではない。そのような状況において、ユーザはしばしば、日本の徳島市所在のＪｕｓｔＳｙｓｔｅｍ　Ｃｏｒｐ．によって製造されたＩｃｈｉｔａｒｏ（登録商標）などのようなワードプロセッサのソフトを使用し、ｒｏｍａｊｉ（日本語における音声的なローマ字（Ｒｏｍａｎ　ａｌｐｈａｂｅｔ）の表現）で書かれたテキストを、カタカナ、ひらがな、および漢字に変換することができる。ワードプロセッサソフトを使用し、ユーザはローマ字でクエリをタイピングし、次いで、ワードプロセッサのスクリーンから変換されたテキストを、ブラウザの検索ボックスへとカットアンドペーストする。このアプローチの不利な点は、相対的に遅く、面倒であり得、ユーザがワードプロセッサのコピーにアクセスすることが要求されるゆえ、コストの制約やメモリの制約などのために、ふさわしいとはいえない。

【０００７】

それゆえ、曖昧な検索クエリに応じて、適切な検索結果を提供する方法および装置が必要とされるのである。

【発明の開示】

【課題を解決するための手段】

【０００８】

具体化され、ここで広く記載される本発明と合致する方法および装置は、曖昧な検索クエリに応じた適切な検索結果を提供する。本発明と合致し、そのような方法は、ユーザからの一連の曖昧な情報構成要素を受け取ることを含む。その方法は、曖昧な情報構成要素を、より曖昧でない情報にマッピングする、マッピング情報を含む。このマッピング情報は、一連の曖昧な情報構成要素を、一つ以上の対応する一連の、より曖昧でない情報構成要素に変換するために使用される。一つ以上のこれらの一連の、より曖昧でない情報構成要素は検索エンジンへの入力として提供される。その検索結果は検索エンジンから得られ

、ユーザに提示される。

【０００９】

付け加えて、システムおよび方法は、検索される文書の少なくとも一部の文字セットまたは言語とは異なる文字セットにて表されたクエリを用いて検索を実行することが開示される。本発明の実施形態により、ユーザは、標準の入力デバイス（例えば、ＡＳＣＩＩキーボード）を用いてクエリをタイプすることができ、クエリをサーバにおいて適切な形式に変換させることができ（たとえば、ローマ字で書かれたクエリをカタカナ、ひらがな、および／または漢字に変換する）、ならびに、変換された形式に基づいて、検索結果を受け取ることができる。

【００１０】

本発明は、プロセス、装置、システム、デバイス、方法、または、コンピュータ可読格納媒体、搬送波、またはコンピュータネットワークなどのコンピュータ可読媒体を含み、多様な方法においてインプリメントされ得ることは理解されるべきであり、プログラムの命令は、光学式または電気の通信線を介して送信される。いくつかの発明の実施形態は以下に記載される。

【００１１】

一実施形態において、方法は、クエリの言葉を、一つの言語および／または文字セットから別のものへと、自動的に変換することが記載される。所定のクエリの言葉を含むアンカーテキストの第１のセットが識別され、それはアンカーテキストが提示する文書（例えばウェブページなど）のセットである。次いで、第２のフォーマットで書かれ、同じ文書のセットを提示するアンカーテキストの第２のセットが識別される。アンカーテキストの第２のセットは、次いで、分析され、第１のフォーマットにおける所定のクエリの言葉の表示が、第２のフォーマットにおける所定のクエリの言葉の表示に対応する確率を得る。

【００１２】

別の実施形態において、確率辞書が作成され、第１のフォーマット（例えば、言語および／または文字セット）で書かれた言葉を、第２のフォーマット（例えば、別の言語および／または文字セット）にマッピングする。確率辞書は、第１のフォーマットで書かれたクエリを第２のフォーマットに変換するために使用される。変換されたクエリは、次いで、検索を実行するために使用され、その結果は、ユーザに戻される。一部の実施形態において、検索結果を用いたユーザの相互作用は、監視され得、確率辞書における確率を更新するために使用される。また、一部の実施形態において、クエリ自体は、検索に先立って、代替的な言語および／または文字セットのマッピングを含むように拡張され得る。

【００１３】

さらなる別の実施形態において、確率辞書を作成する方法が記載される。確率辞書は、第１のフォーマットにおける言葉を第２のフォーマットに変換するために使用され得る。辞書は、アンカーテキストまたはその言葉を含む他のデータを識別することによって、好ましくは言葉毎に作成される。次に、アンカーテキストまたは他のデータに連係される（ａｌｉｇｎｅｄ　ｗｉｔｈ）データは分析され、第１のフォーマットにおける所定の言葉が、第２のフォーマットにおける一つ以上の言葉にマッピングされる確率を決定する。

【００１４】

さらなる別の実施形態において、第１の言語または文字セットに提供されたクエリは、一つ以上のクエリの言葉を含み、第１の言語または文字セットで書かれたアンカーテキストと、第１のアンカーテキストに対応し、第２の言語または文字セットで書かれたアンカーテキストとを比較することによって、第２の言語または文字セットに変換される。

【００１５】

別の実施形態において、コンピュータプログラム製品は、第１のフォーマットで書かれた言葉を第２のフォーマットに変換するために提供される。コンピュータプログラム製品は、コンピュータシステムに、連係されたアンカーテキストを識別させ、第１のフォーマットにおける所定の言葉の表示が、第２のフォーマットにおける一つ以上の言葉に対応する確率を決定させるように動作可能である。

【００１６】
別の実施形態において、方法は、曖昧なクエリを用いて検索を実行するために提供される。ユーザが第１のフォーマットにおいてクエリを入力する場合、それは、第２のフォーマットで書かれた一つ以上の変形の一群に変換される。次いで、検索は、変換された変形を用いて実行され、応答の情報は、ユーザに戻される。例えば、第１のフォーマットは、電話キーパッドを用いて入力された一連の数を含み得、第２のフォーマットは、英数字のテキスト（例えば、英語、ローマ字、ｒｏｍａｊａ、ピンインなど）を含み得る。一部の実施形態において、一つ以上の変形の群は、所定の語彙に現れない、および／または、所定の低い確率の文字の組み合わせを含む、変換された変形を除去することによって選択される。一部の実施形態において、確率辞書は、検索が実行される前に、一つ以上の変形の群を、第３のフォーマットに変換する。例えば、確率辞書は、ローマ字、ｒｏｍａｊａ、またはピンインの一つ以上の変形の群を、漢字、カタカナ、ひらがな、ハングル、ハンジャ、または伝統的な中国文字（ｔｒａｄｉｔｉｏｎａｌ　Ｃｈｉｎｅｓｅ　ｃｈａｒａｃｔｅｒ）に変換するために使用され得、検索は、次いで、変換された変形を用いて実行され得る。

【００１７】
本発明のこれらおよび他の特徴および利点は、以下の詳細な記載、ならびに、本発明の原理の例によって例示された、添付された図面に、さらに詳細に提示されている。

【発明を実施するための最良の形態】

【００１８】
添付された図面は、この明細書にて援用され、その一部として構成され、本発明の実施形態を例示し、記載とともに、本発明の利点および原理を説明するのに役立つ。

【００１９】
添付された図面にて例示される本発明の実施形態を詳細に参照する。同様の数字は、図面や以下に続く記載を通して、同様の部分を示す。以下に続く記載は、当業者が本発明を利用することができるように提示される。特定の実施形態および応用の記載は例としてのみ提供されるのであり、様々な修正は当業者にとって容易に明白である。例えば、多くの例がインターネットのウェブページに記載されているが、本発明の実施形態は、本、新聞、雑誌などの文章および／または情報の他のタイプを検索するために使用され得る。同様に、例示のために、日本語のテキストをローマ字からカタカナ、ひらがなおよび／または漢字へと変換されることが記載されるが、当業者に明らかなように、本発明のシステムおよび方法は、任意の適切な変換へと応用され得る。例えば、限定なしに、本発明の実施形態は、一部の他のフォーマット（例えば、ピンインやローマ字など）において受け取られるクエリに基づき、伝統的な漢字、または韓国のハングル文字またはハンジャ文字にて書かれたテキストを検索するために用いられ得る。ここで記載される一般的な原理は、本発明の趣旨および範囲から逸脱することなく他の実施形態および応用に適用され得る。したがって、本発明は、ここで開示される原理および特徴に合致する多数の代替、修正、および均等物を含み、最も広い範囲に従うものである。明瞭さのために、本発明に関連する領域で既知である技術的事項に関する詳細は、本発明を不必要に曖昧にしないように、詳細に記載されていない。

【００２０】
Ａ．概説
本発明に合致する方法および装置により、ユーザは、曖昧な検索クエリを提出し、場合によっては明確にされた検索結果を受け取ることが可能である。一実施形態において、標準の電話のキーパッドのユーザから受け取る一連の数が、場合によってはそれに対応する英数字のシーケンスのセットに変換される。これらの対応する英数字のシーケンスは、ブール式の「ＯＲ」結果を使用し、従来の検索エンジンへ入力として提供される。この方法において、検索エンジンは、ユーザが興味を持ちそうなものに対する検索結果を制限するのに役立つ。

【００２１】

Ｂ．構成

本発明に合致する方法および装置がインプリメントされ得る、システム１００が図１に示される。システム１００は、ネットワーク１４０を介して、多数のサーバ１２０および１３０に接続される多数のクライアントデバイス１１０を含み得る。ネットワーク１４０は、ローカルエリアネットワーク（ＬＡＮ）、ワイドエリアネットワーク（ＷＡＮ）、公衆交換電話網（ＰＳＴＮ）などのような電話網、インターネット、またはネットワークの組み合わせを含み得る。２つのクライアントデバイス１１０および３つのサーバ１２０および１３０は、単純に、ネットワーク１４０に接続されるように例示される。実際には、ほぼ同数のクライアントデバイスおよびサーバが存在し得る。たま、一部の場合においては、クライアントデバイスは、サーバ機能を実行し得、サーバはクライアントデバイス機能を実行し得る。

【００２２】

クライアントデバイス１１０は、メインフレーム、ミニコンピュータ、パーソナルコンピュータ、ラップトップ、ＰＤＡ（携帯情報端末）などのようなデバイスを含み得、ネットワーク１４０に接続可能である。クライアントデバイス１１０は、ネットワーク１４０を介してデータを送信し、あるいは、有線、無線、または光学式の接続を介してネットワーク１４０からデータを受信する。

【００２３】

図２は、本発明と合致する例示的なクライアントデバイス１１０を図示する。クライアントデバイス１１０は、バス２１０、プロセッサ２２０、メインメモリ２３０、読み出し専用メモリ（ＲＯＭ）２４０、記憶装置２５０、入力デバイス２６０、出力デバイス２７０、および通信インターフェース２８０を含み得る。

【００２４】

バス２１０は、一つ以上の従来のバスを含み得、クライアントデバイス１１０の間で通信を可能にする。プロセッサ２２０は、従来のタイプのプロセッサまたはミニコンピュータを含み得、命令を解釈し実行する。メインメモリ２３０はランダムアクセスメモリ（ＲＡＭ）または他のタイプのダイナミック記憶装置を含み得、プロセッサ２２０による実行のための情報および命令を格納する。ＲＯＭ２４０は、従来のＲＯＭデバイスまたは他のタイプのスタティック記憶装置を含み得、プロセッサ２２０が使用するためのスタティックな情報および命令を格納する。記憶装置２５０は磁気および／または光学式の記憶媒体、ならびにそれに対応するドライブを含み得る。

【００２５】

入力デバイス２６０は、キーボード、マウス、ペン、音声認識および／または生体認識メカニズムなどのような一つ以上の従来のメカニズムを含み得、それによって、ユーザは、クライアントデバイス１１０へ情報を入力することが可能である。出力デバイス２７０は、一つ以上の従来のメカニズム（ディスプレイ、プリンタ、スピーカなど）を含み得、ユーザに情報を出力する。通信インターフェース２８０は、任意の送受信器のようなメカニズムを含み得、クライアントデバイス１１０が他のデバイスおよび／またはシステムと通信することが可能となる。例えば、通信インターフェース２８０は、ネットワーク１４０などのようなネットワークを介して、別のデバイスまたはシステムと通信するためのメカニズムを含み得る。

【００２６】

以下で詳細に記載するように、本発明と合致するクライアントデバイス１１０は、所定の検索に関連する動作を実行する。クライアントデバイス１１０は、メモリ２３０などのコンピュータ可読媒体に含まれるソフトウェアの命令を実行するプロセッサ２２０に応答した動作を実行し得る。コンピュータ可読媒体は一つ以上のメモリデバイスおよび／または搬送波として定義され得る。ソフトウェアの命令は、データ記憶装置２５０などのような別のコンピュータ可読媒体から、または、通信インターフェース２８０を介して別のデバイスから、メモリ２３０へと読み出され得る。メモリ２３０に含まれるソフトウェアの命令により、プロセッサ２２０は、以下で記載される、検索に関連する動きを実行する。

あるいは、ハードウェアに組み込まれている回路は、ソフトウェアの命令の代わりに、またはソフトウェアの命令と組み合わされて使用され得、本発明と合致する処理をインプリメントし得る。したがって、本発明は、特定のハードウェアに組み込まれている回路とソフトウェアとの任意の組み合わせに限定されるわけではない。

【００２７】

サーバ１２０および１３０は、メインフレーム、ミニコンピュータ、またはパーソナルコンピュータなどの、一つ以上のコンピュータシステムのタイプを含み得、ネットワーク１４０と接続することができ、サーバ１２０および１３０は、クライアントデバイス１１０と通信することができる。代替的な実施において、サーバ１２０および１３０は、一つ以上のクライアントデバイス１１０と直接に接続するメカニズムを含み得る。サーバ１２０および１３０は、ネットワーク１４０を介してデータを送信し得、あるいは、有線、無線、または光学式の接続を介して、ネットワーク１４０からデータを受信し得る。

【００２８】

サーバは、クライアントデバイス１１０に対して、図２を参照し上記されたのと同様な方法にて、構成され得る。本発明と合致する実施において、サーバ１２０は、クライアントデバイス１１０によって使用可能である検索エンジン１２５を含み得る。サーバ１３０は、クライアントデバイス１１０によってアクセス可能である文書（またはウェブページ）を格納し得る。

【００２９】

Ｃ．構成動作

図３は、３つの文書を表す図を示し、サーバ１３０のうちの一つにおける例示として格納され得る。

【００３０】

第１の文書（文書１）は、「ｃａｒ　ｒｅｐａｉｒ」および「ｃａｒ　ｒｅｎｔａｌ」の２つのデータ入力を含み、底に「３」という数がある。第２の文書（文書２）は「ｖｉｄｅｏ　ｒｅｎｔａｌ」というデータ入力を含む。第３の文書（文書３）は、「ｗｉｎｅ」、「ｃｈａｍｐａｇｎｅ」、および「ｂａｒ　ｉｔｅｍｓ」という３つのデータ入力を含み、ならびに、文書２へのリンク（または参照）を含む。

【００３１】

例示を単純にするために、図３に示される文書は、英数字文字列の情報（例えば、「ｃａｒ」、「ｒｅｐａｉｒ」、「ｗｉｎｅ」など）のみを含む。しかしながら、当業者は、他の状況において、文書は、音声的、視聴覚的な情報などといった、他のタイプの情報を含み得ることを理解する。

【００３２】

図４ａは、図３に示された文書に基づき、従来の英数字のインデックスを示す。インデックスの第１の列は英数字のリストを含み、第２の列は、それらの言葉に対応する文書のリストを含む。英数字の「３」などの一部の言葉は、一つの文書（この場合は文書１）に対応する（にある）。「ｒｅｎｔａｌ」などの他の言葉は、多数の文書（この場合は文書１および２）に対応する。

【００３３】

図４ｂは、検索エンジン１２５のような従来の検索エンジンが、図４ａにて示されたインデックスをどのように使用し、英数字の検索クエリに応じた検索結果を提供するのかを示す。英数字のクエリは、任意の従来の技術を用いて生成され得る。図示のために、図４ｂは２つの英数字のクエリ、「ｃａｒ」および「ｗｉｎｅ」を表す。従来のアプローチの下では、検索エンジン１２５は、「ｃａｒ」などの英数字のクエリを受け取り（ステージ４１０）、その英数字のインデックスを用いて、どの文書がそのクエリに対応するのかを決定する（ステージ４２０）。この例において、従来の検索エンジン１２５は、図４ａにて示されたインデックスを用い、「ｃａｒ」が文書１に対応し、検索結果としてユーザに、文書１（またはそれへの参照）を戻す。同様に、従来の検索エンジンは、「ｗｉｎｅ」が文書３に対応し、ユーザに文書３（またはそれへの参照）を戻すことを決定する（ステ

ージ４３０）。

【００３４】

図５ａは、本発明と合致し、個々に図３および図４ａに示された文書およびインデックスに基づき、数字の検索クエリに応じた検索結果を提供する好ましい技術のフロー図を示す。理解し易い例示のために、図５ａは標準の電話端末のマッピングに基づき、数字のクエリを処理する特定の技術を記載する。しかし、当業者は、本発明に合致する他の技術もまた使用され得ることを理解する。

【００３５】

ステージ５１０において、シーケンス「２２７」（数字の構成要素「２」、「２」、および「７」からなる）がユーザから受け取られる。ステージ５２０において、数字の構成要素が文字へとマッピングされる方法についての情報が得られる。ユーザが標準の電話キーパッドから情報を入力したと想定し、このマッピング情報が図５ｂに示される。図５ｂに示されるように、文字「ａ」、「ｂ」、および「ｃ」はそれぞれ、「１」にマッピングされ、文字「ｐ」、「ｑ」、および「ｒ」はそれぞれ、「７」にマッピングされている。

【００３６】

ステージ５３０において、このマッピング情報を用い、シーケンス「２２７」は、その英数字の相当物に変換される。図５ｂに示される情報に基づいて、シーケンス「２２７」に対応する文字の可能な組み合わせは３６通りあり、その組み合わせは、ａａｐ、ｂａｐ、ｃａｐ、ａｂｐ、ｂｂｐ、・・・ｂａｒ・・・ｃａｒ・・・ｃｃｓなどを含む。数字が可能な組み合わせ（例えば「ａａ７」）に含まれる場合、８０通りの可能な組み合わせが存在する。全ての可能な英数字の相当物を生成するよりもむしろ、一部の語彙に基づき、生成された相当物を限定するのが望ましい。例えば、辞書、または以前の検索クエリの検索エンジンログなどに存在する英数字の相当物のみを生成することが望ましい。あるいは、既知の統計的な技術（例えば、所定の言葉が一緒に現れる確率など）を用いることによって英数字の相当物を限定することが望ましい。

【００３７】

ステージ５４０において、これらの英数字の相当物は、論理「ＯＲ」動作を用いて、図４ａおよび図４ｂを参照して記載されたような、従来の検索エンジンへの入力データとして提供される。例えば、検索エンジンへ提供される検索クエリは、「ａａｐ　ＯＲ　ｂａｐ　ＯＲ　ｃａｐ　ＯＲ　ａｂｐ・・・ＯＲ　ｂａｒ・・・ＯＲ　ｃａｒ」であり得る。全ての可能な英数字の相当物が検索クエリに提供され得るが、サブセットは、その代わりに、意図されない相当物を除去する従来の技術を用いて使用され得る。例えば、ユーザは、言葉の使用についての確率的な情報を引き出す技術を用いて、可能な組み合わせのより狭いリストを生成することが可能である。すなわち、ユーザは、「ｑｕ」で始まる組み合わせを含む（好む）が「ｑｔ」で始まる組み合わせを無視することが可能である。

【００３８】

ステージ５５０において、検索結果は検索エンジンから得られる。なぜなら、「ａａｐ」や「ａｂｐ」などの言葉は検索エンジンのインデックスには存在せず、それらは効果的に無視されるからである。実際には、図４ｂに示されたインデックス内に含まれた言葉は「ｃａｒ」および「ｂａｒ」のみであり、戻ってきた検索結果は、文書１および文書３を参照するもののみである。ステージ５６０において、これらの検索結果はユーザに提示される。その検索結果は、検索エンジンによって提供されたのと同じ順序にて提示され得るか、または、ユーザの言語などの検討材料に基づき、記録され得る。ユーザが「ｂａｒ」という言葉を含む文書のみに興味を持っていると想定すると、ユーザは、望んでいない結果（文書３）、および望んだ結果（文書１）を受け取る。ユーザの利点として、検索クエリを編成する３つのキーを押すことが必要とされるのみであるが、これは許容可能な対価であり得る。

【００３９】

図６は、本発明に合致し、個々に図３および図４ａに示された文書およびインデックスに基づき、数字の検索クエリに対応する検索結果を提供する好ましい技術の、別のフロー

図を示す。このフロー図は、受け取られたシーケンスのサイズの増加が、ユーザによって望まれたものへの検索結果をどのように限定するのを助けるのかを例証する。理解し易い例示のために、図６は、標準の電話キーパッドのマッピングに基づいて、数字のクエリを処理する特定の技術を再び記載するが、当業者は、本発明に合致する他の技術が利用され得ることを理解する。

【００４０】

ステージ６１０において、シーケンス「２２７　４８３６７」（数字の構成要素、「２」、「２」、「７」、「４」、「８」、「３」、「６」、「７」からなる）が、ユーザから受け取られる。説明のために、シーケンス「２２７」を「数字ワード」（ｎｕｍｂｅｒ　ｗｏｒｄ）と呼び、全体のシーケンス「２２７　４８３６７」を「数字フレーズ」（ｎｕｍｂｅｒ　ｐｈｒａｓｅ）と呼ぶ。数字ワードの可能な英数字の相当物を「文字ワード」（ｌｅｔｔｅｒ　ｗｏｒｄ）と呼び、数字フレーズの可能な英数字の相当物を「文字フレーズ」（ｌｅｔｔｅｒ　ｐｈｒａｓｅ）と呼ぶ。

【００４１】

ステージ６２０において、数字の構成要素が文字にマッピングされる方法についての情報が得られる。ステージ６３０において、同様のマッピング情報が、図５ｂに示されるように使用されると想定し、数字フレーズ「２２７　４８３６７」は、それに対応する文字フレーズに変換される。図５ｂに示される情報に基づき、シーケンス「２２７　４８３６７」に対応する、１１６６４通りの可能な文字フレーズが存在する。

【００４２】

ステージ６４０において、これらの文字フレーズは、論理「ＯＲ」動作を用いて、図４ａおよび図４ｂを参照して記載された、従来の検索エンジンへの入力データとして提供される。例えば、検索エンジンに提供された検索クエリは、「ａａｐ　ｇｔｄｍｐ‘ＯＲ’ａａｐ　ｈｔｄｍｐ‘・・・ＯＲ’ｂａｒ　ｉｔｅｍｓ‘・・・ＯＲ’ｃａｒ　ｉｔｅｍｓ‘」であり得る。全ての可能な文字フレーズが検索エンジンに提供され得るが、サブセットは、その代わり、意図されていない文字フレーズを除去するために従来の技術を用いて、使用され得る。

【００４３】

ステージ６５０において、検索結果は検索エンジンから得られる。なぜなら、多くの検索エンジンが、ソートされた正確なフレーズを含むそれらの文書を上位にランクさせるように設計されており、文書３は、最上位にランクされた検索結果であるからである（つまり、正確なフレーズ、「ｂａｒ　ｉｔｅｍｓ」を含むからである）。例における文書で、ステージ６２０にて生成された他の文字フレーズのうちの一つを含む文書はない。さらに、多くの検索結果は、フレーズの個々の部分を含む検索結果を減らす（除去する）が、全体のフレーズではない。例えば、文書１は、それが、「ｃａｒ」という文字ワードを含むゆえに減らされ（除去され）、その文字ワードは文字フレーズの第１の部分に対応するが、それは、文字フレーズの第２の部分に対応する任意の文字ワードを含まない。最後に、「ａａｐ　ｈｔｄｍｐ」などの文字フレーズは効果的に無視される。というのは、それらは、検索エンジンのインデックスに存在する文字ワードを含まないからである。

【００４４】

ステージ６６０において、検索結果がユーザに提示される。例において、ユーザに示された第１の結果は文書３であり、それは、ユーザのクエリに最も適切なものである。文書１は、可能な文字フレーズのうちの一つを含まないゆえに、共に除去され得る。この方法において、ユーザは、最も適切な検索結果が提供される。

【００４５】

図５および図６を参照した上記は、数字の情報を受け取り、それを英数字の情報にマッピングすることを参照してなされるが、当業者は、他の実施が本発明と合致して可能であることを理解する。例えば、ユーザによって押されたキーに対応する数のシーケンスを受け取る代わりに、受け取られたシーケンスは、ユーザによって押されたキーに対応する第１の文字からなり得る。つまり、「２２７」を受け取る代わりに、受け取られたシーケン

スは「ａａｐ」であり得る。本発明と合致し、ステージ５３０または６３０において生成された、その相当する文字シーケンスは、「ａａｐ」に対応する他の文字シーケンス（例えば「ｂａｒ」）であり得る。実際には、受け取られたシーケンスは、音声的、視聴覚的、または他の任意の情報構成要素のタイプを含み得る。

【００４６】

シーケンスが受け取られるフォームに関係なく、受け取られたシーケンスは、情報が検索エンジンのインデックスに格納されるフォーマットに対応するシーケンスに変換されるのが、通常は好ましい。例えば、検索エンジンのインデックスが英数字のフォーマットにて記憶される場合、受け取られたシーケンスは英数字のシーケンスに変換されるべきである。

【００４７】

さらに、情報構成要素の受け取られたシーケンスを変換するために使用されるマッピング技術は、ユーザの入力をデバイスによって生成される情報にマッピングするためのユーザのデバイスにて用いられるのと同じ技術であることが通常は望ましい。しかしながら、ユーザの入力に使用されるのとは異なるマッピング技術が使用されるのが好ましい場合もあり得る。

【００４８】

また、本発明の実施形態により、ユーザは、ターゲットではない言語のキーボードを用いて入力された検索を実行することを可能にする。例えば、日本語のテキストを含むウェブページは漢字で書かれる一方、そのページを検索しようとするユーザは、ローマ字のアルファベットに基づいて標準のＡＳＣＩＩキーボード（または携帯電話機）にアクセスを有するのみである。

【００４９】

図７は、そのような検索を実行する方法を例示する。図７において示されるように、ユーザは、標準の入力デバイス（例えば、ＡＳＣＩＩキーボード、携帯電話機、など）を使用してクエリをタイプし、そのクエリを検索エンジンに送る。そのクエリは、それに対応する文書の一部が書かれている（例えば漢字）文字セットとは異なる文字セット（例えばローマ字）で書かれ得る。検索エンジンは、クエリを受信し（ブロック７０２）、それを適切な形式に変換し（ブロック７０４）、従来の検索技術などを使用して、変換されたクエリに対応して、文書の検索を実行する（ブロック７０６）。次いで検索エンジンは、対応する文書のリスト（および／または、文書のコピー）を、ユーザに戻す（ブロック７０８）。例えば、図６と関連して上記されたものと同様の方法にて、結果はユーザに戻され得る。

【００５０】

図７に示されるように、ユーザのクエリは、好ましくは、クライアントとは逆の、検索エンジンサーバ側にて変換され、変換を実行するための、特殊な目的のためのソフトウェアを得る必要から、ユーザを解放させる。しかしながら、他の実施形態において、変換の一部または全ては、クライアント側にて実行され得ることは理解される。付け加えて、一部の実施形態において、クエリは、電話機のキーパッドなどのようなデバイスを使用して入力され得る。そのような実施形態において、初期の数字のクエリは、図５および図６に関連する上記されたマッピング技術を用いて、英数字の形式（例えばローマ字）に変換され得、例えば、低い確率のマッピング（例えば、ローマ字においては生じない文字の組み合わせなどを含むマッピング）を除去するための、語彙および／または確率的な技術の応用を含む。いったん、クエリの英数字の変換が得られると、図７に示されるステップの残りが実行され得る（つまり、７０４、７０６、および７０８）。

【００５１】

一つの文字セットまたは言語から別のものへの変換（つまり、図７におけるブロック７０４）は、様々な方法にて実行され得る。一つの技術は、クエリにおけるそれぞれの言葉を、ターゲットの言語または文字セットにおいて対応する言葉にマッピングするために、言葉の意味または変換の従来のスタティックの辞書を使用する。しかしながら、このアプ

ローチの問題は、しばしば不正確な結果を生じることである。というのは、言葉はしばしば曖昧であり、クエリはしばしば、短すぎて、この曖昧さを解消する十分な手掛かりを提供しない。例えば、「ｂａｎｋ」という言葉は、川の土手、金融機関、飛行機による演習、などを意味し得、理論上、正確に変換することは困難である。付け加えて、辞書が相対的に、大きくなく、および／または頻繁に更新されていない場合、滅多に使用しない言葉、スラング、イディオム、適切な名前などの、検索エンジンが出くわし得る全ての言葉の入力が含まれ得ない。

【００５２】

本発明の実施形態はまた、一つの言語または文字セット（例えばＡＳＣＩＩ）から別のもの（例えば漢字）へ、クエリの言葉を変換するために、確率辞書（ｐｒｏｂａｂｉｌｉｓｔｉｃ　ｄｉｃｔｉｏｎａｒｙ）を使用して、一部または全ての問題を克服または改善するために使用され得る。好ましい実施形態において、確率辞書は、一つの言葉のセットを別の言葉のセットへとマッピングし、確率をそのマッピングそれぞれに関連付ける。便宜上、「言葉（ｔｅｒｍ）」または「トークン」は、言葉（ｗｏｒｄｓ）、フレーズ、および／または、スペースを含み得る一つ以上の文字のシーケンスを参照する。

【００５３】

図８は、上記されたような確率辞書８００の例を示す。図８に示された例示的な確率辞書８００は、ｒｏｍａｊｉ（日本語のローマ字のアルファベットの表示）で書かれた言葉を、漢字（ローマ字ではない、表意文字ベースの日本語の文字セット）で書かれた言葉にマッピングする。説明を容易にするために、図８は、ローマ字の言葉を、「＜ｔｅｒｍ＞$_{ｒｏｍａｊｉ}$」とし、漢字の言葉を「＜ｔｅｒｍ＞$_{ｋａｎｊｉ}$」とする。漢字辞典に対する実際のローマ字において、実際のローマ字および漢字の言葉は、図８にて示される英語変換よりも、使用されることが理解される。したがって、図８は、本発明の実施形態の説明を容易にするために提供されるのであり、日本語のテキストの実際の文字および意味を例示しているのではない。

【００５４】

辞書８００は、様々なローマ字の言葉８０２のための８０８、８１０、８１２、８１４のデータ入力を含む。辞書はまた、漢字８０４におけるそれぞれの言葉の表示を含み、それぞれの表示が正しい場合の、対応する確率８０６に沿っている。例えば、ローマ字の言葉「ｂａｎｋ」は、「ｓｔｅｅｐ　ｓｌｏｐｅ」を意味する言葉に、０．３の確率でマッピングされ得、「ｆｉｎａｎｃｉａｌ　ｉｎｓｔｉｔｕｔｉｏｎ」を意味する言葉には、０．４の確率でマッピングされ得、「ａｉｒｐｌａｎｅ　ｍａｎｅｕｖｅｒ」には０．２の確率でマッピングされ得る。０．１の確率では、その言葉は「その他」にマッピングされ得、それは単に、辞書に存在し得ない言葉に、それぞれの言葉をマッピングさせる、包括的な方法である。

【００５５】

再び、図８に示された例は、第１の文字セットまたは言語における所定の言葉（例えば、「ｂａｎｋ」など）が、別の文字セットまたは言語において２つ以上の言葉にマッピングされ得ることを例示するように構成されることが理解される。しかしながら、当業者が理解するように、明瞭さのために、図８における特定の例は、英語の言葉や意味を使用する原理を例示しており、「ｂａｎｋ」などの言葉の実際のローマ字表示は、例えば、その英語の相当物と同じような方法にて曖昧ではあり得ない（例えば、ｆｉｎａｎｃｉａｌ　ｉｎｓｔｉｔｕｔｉｏｎとａｉｒｐｌａｎｅ　ｍａｎｅｕｖｅｒとの言葉の間で、ローマ字に曖昧さは存在し得ない）。また、理解すべきことは、説明を容易にするために、図８に示される辞書は、他の点においても単純化されている。例えば、実施の確率辞書は、それぞれの言葉の、さらに多くのマッピングを含み得、または、所定の確率閾値を超過するマッピングを含み得る。

【００５６】

本発明の好ましい実施形態は、そのような確率辞書を用い、一つの言語および／または文字セットにおいて表現されたクエリを、別の言語および／または文字セットに変換し、

それにより、ユーザが、元々のクエリとは異なる文字セットおよび／または言語にて書かれた文書を発見することを可能にする。例えば、ユーザがローマ字で「ｃａｒｓ」というクエリを入力した場合、確率辞書は、「ｃａｒｓ」というローマ字の言葉を、「ｃａｒｓ」という漢字の言葉へとマッピングするために使用され得る。この方法において、クエリの文字セット（例えばローマ字）とマッチングする文書の文字セット（例えば漢字）とが同じではない場合でさえも、ユーザは、それらのクエリに関連する文書を発見することができる。この特定の例において、クエリの実際の言語は変化せず（ローマ字および漢字は日本語を表現するために使用される）、文字エンコードのみが変化することに注意されたい。

【００５７】

別の例として、ＡＳＣＩＩの英語における「ｔｉｒｅｄ」という言葉は、Ｌａｔｉｎ１の文字エンコードを用いたドイツ語における「ｍｕｅｄｅ」という言葉にマッピングされ得る。というのは、ウムラウトｕという文字は、ＡＳＣＩＩに存在しないからである。この例において、辞書は、辞書は他の言語に変換され（英語からドイツ語）、他の文字エンコード（ＡＳＣＩＩからＬａｔｉｎ１）へと変換されることに注意されたい。

【００５８】

好ましい実施形態において、上記されたマッピング辞書は、自動的な方法において構築され、統計的な技術に関連して、ウェブ上で利用可能な情報を用いる。好ましい実施形態は、正確な変換に達するために、異なる言語および／または文字セットにて書かれたアンカーテキストなどのような、パラレルで連係されたバイリンガルのコーパスを用いる。このデータを用いて、好ましい実施形態は、言葉のマッピングの辞書を構築することが可能である。これは、例えば、単に言語S_i（ソース言語）が、連係されたテキストの対（例えばアンカー、文、など）におけるトークンT_j（ターゲット言語）と同時に生じる回数を数えることによってなされ得る。しかしながら、任意の適切な技術が用いられ得ることは理解されたい。十分に広く正確に連係されたセットが存在しない場合において、この方法は、相対的に曖昧な多対多のマッピングを生成し得る。したがって、例えば、S_1は、一部の確率を用いて、T_2、T_3、T_7およびT_8にマッピングされ得ることが決定され得る。しかしながら、これは、以下で詳細に記載するように、許容可能であり、一部の実施形態において、追加的な改良が、それぞれのマッピングの個々の可能性（例えば、以前のユーザのクエリ、結果ページ上のアイテムのユーザ選択および／またはそのようなものを調べることによって）を増加させるためになされ得る。

【００５９】

図９は、確率辞書を構築するための、パラレルアンカーテキストの使用を示す。アンカーテキストは、ウェブページ間（または、所定のウェブページ内の位置）のハイパーリンクに関連付けられたテキストを含む。例えば、ハイパーテキストマークアップ言語（ＨＴＭＬ）において、「＜Ａ　ｈｒｅｆ＝”ｈｔｔｐ：／／ｗｗｗ．ａｂｃ．ｃｏｍ”＞Ｂａｎｋｓ　ａｎｄ　Ｓａｖｉｎｇｓ　ａｎｄ　Ｌｏａｎｓ＜／Ａ＞」というコマンドは、「Ｂａｎｋｓ　ａｎｄ　Ｓａｖｉｎｇｓ　ａｎｄ　Ｌｏａｎｓ」というテキストを、ｈｔｔｐ：／／ｗｗｗ．ａｂｃ．ｃｏｍのウェブページを提示するハイパーリンクとして表示させる。「Ｂａｎｋｓ　ａｎｄ　Ｓａｖｉｎｇｓ　ａｎｄ　Ｌｏａｎｓ」というテキストは、アンカーテキストと呼ばれ、通常は、提示されるウェブページ（例えば、ｗｗｗ．ａｂｃ．ｃｏｍ）の短い記載を提供する。実際は、アンカーテキストは、しばしば、そのページ自体よりも、より正確なウェブページの記載を提供し、提示するウェブページの性質を決定するのに、特に有用であり得る。付け加えて、アンカーテキストにおける言葉の使用および配布は、しばしば、趣旨および長さにおいて、ユーザのクエリにて見出されるものに近い。また、所定のページを提示するアンカーの多くは、同じか、または高度に類似したテキストである場合がある。例えば、ｗｗｗ．ｇｏｏｇｌｅ．ｃｏｍを提示するアンカーは、単に、「Ｇｏｏｇｌｅ」であるか、または、他のテキストとともに、この言葉を少なくとも使用する。したがって、例えばカタカナなど、ｗｗｗ．ｇｏｏｇｌｅ．ｃｏｍを提示する全てのアンカーを検証することで、「Ｇｏｏｇｌｅ」のカタカナ変換は、最も高

い頻度で現れる言葉を単に探すことによって、相対的に高い信頼度でもって推測され得る（可能であれば、単なる「ここをクリック」といったような、所定の低い情報内容を除去した後で）。本発明の好ましい実施形態は、正確な変換を提供するために、アンカーテキストのこれらの特性の利点を持つ。

【００６０】

図９を参照すると、第１の文字セット（例えばＡＳＣＩＩ）に書かれた言葉を含むクエリを受け取ると（ブロック９０２）、サーバは、言葉にあるアンカーテキストのセットを検証し得る（ブロック９０４）。例えば、サーバは、その言葉を含むそれらのアンカーを識別するために、全ての既知のアンカーのインデックスを検証し得る。次に、それらのアンカーが提示するウェブページは識別され（ブロック９０６）、アンカーは、それらのページを提示するターゲット言語またはターゲット文字セット（例えば、ひらがな、カタカナ、および／または漢字）で書かれる（ブロック９０８）。システムはここで、２つの文書のセットを有する（そこでは、アンカーテキストは、文書のフォームと見なされる）。一つの文書のセット（例えば、もともとのＡＳＣＩＩのクエリを含むアンカー）におけるクエリの言葉の分配は、次いで、他の文書セット（例えば、パラレルアンカー）における変換されたフレーズの最も適当な候補を識別するために使用される。統計は、アンカーテキストが現れる頻度に関して計算され得、これらの統計は、アンカーテキストに見出される言葉の相対的な頻度または確率が、もともとのクエリの正しい変換であるかどうかを決定するために使用され得る（ブロック９１０）。多数の言葉を有するクエリに対して、上記のプロセスは、それぞれの言葉に対して繰り返され得、または、全体のクエリは、単一の言葉として扱われ得、あるいは、一部の他の適切な言葉の群が使用され得る。例えば、クエリが「ｂｉｇ　ｈｏｕｓｅｓ」である場合、可能な変換の辞書は、そのフレーズを含む、連係されたアンカーテキストを見出すことによって構築され得る。同様に、クエリが３つ以上の言葉を含む場合、適切なマッピングを決定するための経験は、クエリの言葉の適切なサブセットを取り上げ、それらの言葉の結果を生成することによって構築され得る。

【００６１】

図９に示される方法において変換を実行する利点は、変換システムが一つの言語または文字セットにおける言葉とターゲットセットにおけるそれらとの間のマッピングの予備的知識を必要としない。その代わり、マッピングは、統計的分析を実行するために利用可能であるデータの本体に基づいて、ダイナミックに決定され得る。したがって、例えば、従来のスタティックな辞書を維持する労力または費用（例えば、言語的分析および調査）を負うことなしに、スラング、イディオム、適切な名前などに対する正確な変換を発見することが可能である。

【００６２】

前述の変換の例示的な実施形態は、ここで、図１０～図１２と関連して記載される。この例において、ユーザは「ｈｏｕｓｅ」というクエリの言葉を入力し、スペイン語で書かれた検索結果（または、単に、クエリの言葉の変換）を得ることを望んでいると想定する。サーバは、英語の「ｈｏｕｓｅ」を、スペイン語の相当物に変換することを企てる。

【００６３】

図１０を参照すると、様々なウェブページ９５９、９６１、９６３、９６５が、アンカーテキスト９６０、９６２、９６４、９６６を介して、ページ９７２および９７４にリンクされる。一部のページおよびそれらに関連するアンカーテキストは、英語で書かれており（つまり、ページ９５９ａ～９５９ｅおよび９６３ａ～９６３ｔ）、一部はスペイン語で書かれている（ページ９６１ａ～９６１ｅおよび９６５ａ～９６５ｊ）。サーバは、第１に、「ｈｏｕｓｅ」という言葉を使用する全てのアンカーの位置を突き止める。これらのアンカーは、例えば、サーバにおいて格納されたアンカーテキストのインデックスを検索することによって、位置が突き止められ得る。そのようなインデックスを使用して、サーバは第１に５つのアンカー９６０を見出し得、それぞれが「ｂｉｇ　ｈｏｕｓｅ」というフレーズを使用し、ウェブページ９７２を提示する。サーバはまた、次に、ページ９７

２を提示する５つのターゲット言語（例えばスペイン語）のアンカー９６２が存在することを決定する。図１０に示される例において、これらのアンカーは「ｃａｓａ ｇｒａｎｄｅ」というテキストを含む。同じページ（アンカー９６０およびアンカー９６２など）、またはそれに対して所定の関係を有するページに提示されるアンカーは、「連係（ａｌｉｇｎｅｄ）」されているといい、そこでは、より一般的な意味では、配置が通常、連係されたアイテムの相当物（またはほぼ相当物）を意味する。

【００６４】

図１１Ａは、それぞれのターゲット言語の言葉は、ターゲット言語のアンカー９６２に現れる頻度を示す。図１１Ａに示されるように、「ｃａｓａ」および「ｇｒａｎｄｅ」はそれぞれ、５回現れる（つまり、それぞれのアンカー９６２に一度）。したがって、ターゲットアンカー９６２に現れる、トータルで１０の言葉（つまり、５つのアンカーのそれぞれにおいて、アンカー毎に２つの言葉）から、「ｃａｓａ」は半分を占め、「ｇｒａｎｄｅ」はもう半分を占める。したがって、図１１Ａで示されるように、この時点で、「ｈｏｕｓｅ」という言葉は、「ｃａｓａ」および「ｇｒａｎｄｅ」の両方の言葉が等しい頻度で現れるゆえに、等しい確率で、「ｃａｓａ」または「ｇｒａｎｄｅ」のいずれかにマッピングされ得る。

【００６５】

しかしながら、図１０にて示されるように、システムはまた、「ｈｏｕｓｅ」という言葉を含む２０個の英語のアンカー９６４を見出し、ページ９７４に提示し、ならびに、「ｃａｓａ」という言葉を含む１０個のスペイン語のアンカー９６６を見出し、ページ９７４を提示する。図１１Ｂにて示されるように、「ｈｏｕｓｅ」という言葉は、「ｃａｓａ」という言葉に、０．７５の確率（つまり、１５／２０）でマッピングされ、「ｇｒａｎｄｅ」という言葉に、０．２５の確率（つまり５／２０）でマッピングされる。これらの確率は、そのターゲット言語のアンカーにおけるそれぞれの言葉の出現回数のトータル（「ｃａｓａ」の場合は１５）を、ターゲット言語のアンカーにおける言葉のトータル数（重複を含む）によって、単に割ることによって計算される（つまり、２０の言葉とは、アンカー９６２に含まれる１０、および９６４に含まれる１０）。代替的には、または追加的には、他の技術が、所定の変換またはマッピングの確率を計算および／または改良するために使用され得る。例えば、当業者が理解するように、任意の様々な既知の技術は、ベイズ法（Ｂａｙｅｓｉａｎ ｍｅｔｈｏｄｓ）、ヒストグラムスムージング（ｈｉｓｔｏｇｒａｍ ｓｍｏｏｔｈｉｎｇ）、カーネルスムージング（ｋｅｒｎｅｌ ｓｍｏｏｔｈｉｎｇ）、縮小推定量（ｓｈｒｉｎｋａｇｅ ｅｓｔｉｍａｔｏｒｓ）、および／または他の推定方法などの確率推定の分散エラーを減少させるために使用され得る。

【００６６】

さらなるアンカーテキストが利用可能である場合、確率は、一層さらに改良され得る。例えば、最終確率分配は、図１２にて示されたものと同様であり得、「ｈｏｕｓｅ」は、相対的に高確率で、「ｃａｓａ」、および、それに接尾語が付いた形式「ｃａｓｉｔａ」にマッピングされ、幾分か低い確率で、「ｃａｓｉｎｏ」および

【００６７】

【数１】

“mansión”

（スペイン語でｍａｎｓｉｏｎに相当する）にマッピングされ、僅かな確率で、「ｇｒａｎｄｅ」にマッピングされる。したがって、正確な変換、およびほぼ同意語の識別は、変換される言語および／または文字セットの知識なしで得られ得る。

【００６８】

クエリの言葉を変換すると、サーバは、変換を使用して検索を実行し得る。例えば、ユーザは「ｈｏｔｅｌｓ ｉｎ Ｋｙｏｔｏ」に対するローマ字のクエリを入力する場合、上記された技術は、サーバに、カタカナ、ひらがな、および漢字のクエリのフォームを推測させることができ、それらのクエリを使用して検索を実行させることができ、ならびに

、適切なユーザインターフェース内で、ユーザへのそれらのクエリのフォームのそれぞれに対する、組み合わされた結果を提示させることができる。

【００６９】

図１０～図１２に関連して記載された例は、例示のために提供されるのであり、限定のためではなく、多くの変更がそこに表された方法論に対してなされ得ることは理解されるべきである。例えば、異なる統計的な技術が一定の確率に達するために使用され得、および／または、修正は、上記された基本的な技術に対してなされ得る。同様に、上記された変換技術は、単に、ユーザによって入力された言葉またはフレーズの変換を実行するために使用され得、関連するインターネット検索を実行し、または確率辞書を作成するために使用される必要はない。付け加えて、先行する例は、ユーザのクエリの受け取りのアプリケーションとに生じるものとして、変換プロセスを記載するが、他の実施形態においては、マッピングプロセスは、ユーザのクエリが受け取られる前に実行され得ることは理解されるべきである。そのような計算前のマッピングは、図８に記載されたような辞書に格納され得、次いで、それらが受け取られたときに、ユーザのクエリを変換するように適合される。最後に、連係されたアンカーテキストとは異なるテキストは、変換を実行するために使用され得ることは理解されるべきである。例えば、連係された文章または他のデータは、同様な方法にて使用され得る。多くの国では、一つ以上の公式言語または認められた言語が存在しており、新聞や定期刊行物はしばしばこれらの言語のそれぞれにて書かれた同じ記事を含む。これらのパラレルな変換は、言葉の変換の確率辞書を準備するために、前記されたアンカーテキストと同様の方法において使用され得る。

【００７０】

したがって、好ましい実施形態により、有利にも、ユーザは、従来の方法において、検索クエリおよび／または変換要求を入力することができ、正確で自動の変換および検索を提供する。一部の実施形態において、追加の改良が上記された基本的なモデルに対してなされ得る。例えば、一部の実施形態において、優先（加重）は、もともとのクエリおよび／または他の連係されたアンカーにおける言葉の数と同様の言葉の数を含むアンカーに与えられ得る。例えば、図１０に示されるシステムにおいて、優先は、ページ９７４に提示されるアンカーに与えられ得る。というのは、もともとのクエリと同様に、それらは、それぞれ単一の言葉を含むからである。同様に、「ｌａ　ｃａｓａ　ｇｒａｎｄｅ」のテキストを含むアンカーがまたページ９７２に提示された場合、その加重は、適切な要因によって軽減され得、というのは、それが連係された他のアンカーよりも多くの言葉（例えば３つ）を含むからである。そのような加重スキームは、適切な要因によってこれらのアンカーの言葉と関連される頻度を増加させることによって、図１１Ｂに示される確率計算に反映され得る。

【００７１】

上記された変換プロセスはまた、検索それ自体の効率を改善するために利用され得る。例えば、確率辞書は、様々な変換およびもともとのクエリの言葉の同意語などを含む、進行中のクエリを拡張するために使用され得る。文書検索に先立つユーザのクエリを拡張することによって、同じ「概念」の同時検索が実行され得、それにより、検索結果は、ユーザが探しているものを含む可能性を増加させる。代替的には、または追加的には、確率辞書は、文書の言葉の拡張を提供することによって、通常の文書インデックス付けのプロセスを補うために使用され得る。例えば、文書にて見出される言葉は、確率辞書からの変換を用いて、文書のインデックスにおいて補われ得、その文書は、もともとの文書にて見出された同じ言葉を正確に使用しない検索によってさえも、位置を突き止められる可能性を増加させる。

【００７２】

上記された変換技術を使用する場合に生じ得る問題は、データの希薄（例えば、「ｃａｓａ」を「ｈｏｕｓｅ」にマップすることを最終的に決定するには十分なアンカーがない）、または、多様性の欠如（全てのアンカーが同じものを示す）などのためであり、システムは、十分に正確な確率のマッピングに達することが不可能であり得る。したがって、

一部の実施形態において、確率マッピングは、ユーザの行動を検証することでさらに改良され得る。いくつかの例示的な技術が以下に記載される。

【００７３】

例えば、再度、サーバが「ｈｏｕｓｅ」に対する変換を得ることを望んでいる場合を想定する。しかしながら、見出され得るアンカーテキストが、「ｂｉｇ　ｈｏｕｓｅ」というフレーズ、または「ｃａｓａ　ｇｒａｎｄｅ」というフレーズのいずれかを含むことを想定する。そのアンカーテキストにおける多様性の欠如のために、確率辞書は以下のマッピングに達し得る。

ｈｏｕｓｅ→ｃａｓａ、０．５の確率
ｈｏｕｓｅ→ｇｒａｎｄｅ、０．５の確率

ｂｉｇ→ｃａｓａ、０．５の確率
ｂｉｇ→ｇｒａｎｄｅ、０．５の確率

ｇｒａｎｄｅ→ｈｏｕｓｅ、０．５の確率
ｇｒａｎｄｅ→ｂｉｇ、０．５の確率

ｃａｓａ→ｈｏｕｓｅ、０．５の確率
ｃａｓａ→ｂｉｇ、０．５の確率

ここでユーザが「ｃａｓａ」という言葉を用いて検索エンジンにクエリすると想定する。この時点で、検索エンジンは、「ｃａｓａ」という言葉を含むページを返答し得、また、「ｈｏｕｓｅ」という言葉をちょうど含むＮ結果と、「ｂｉｇ」という言葉をちょうど含むＭ結果とを合わせる。実際には、ＮおよびＭは、マッピングに内在する確率を考慮するように調整され得、その結果、相対的に見込みのないマッピングは、より少ない結果が表示されることになる。ユーザが、「ｂｉｇ」という言葉を含む結果をクリックするよりも、「ｈｏｕｓｅ」という言葉を含む結果を１０倍、クリックした場合、マッピングの確率は、例えば、以下のように調整される。

ｈｏｕｓｅ→ｃａｓａ、０．９の確率
ｈｏｕｓｅ→ｇｒａｎｄｅ、０．１の確率

ｂｉｇ→ｃａｓａ、０．１の確率
ｂｉｇ→ｇｒａｎｄｅ、０．９の確率

ｇｒａｎｄｅ→ｈｏｕｓｅ、０．１の確率
ｇｒａｎｄｅ→ｂｉｇ、０．９の確率

ｃａｓａ→ｈｏｕｓｅ、０．９の確率
ｃａｓａ→ｂｉｇ、０．１の確率

実際の数は、クリックが考慮されるユーザの数、その言葉の両方を含むページのクリックの回数、結果のセットの中の、当該の言葉を含む結果の置換、および／またはそれらのようなものなど、他の様々な要因に依存し得ることに注意されたい。この例（つまり、０．１および０．９）に与えられた調整された確率は、例示の目的のためであることは理解されたい。当業者は、上記されたものなどのユーザフィードバックに与えられた実際の加重は、任意の適切な方法にてインプリメントされ得ることは理解する。

【００７４】

また、前述の例は、ユーザフィードバックの使用の例を容易にするように簡易化されていることに注意されたい。例えば、一部のシステムにおいて、所定の変換を実行することを補助するために、他の変換から得られた情報を利用することが可能である。例えば、今

提示された例において、「ｈｏｕｓｅ」という言葉が「ｂｉｇ　ｈｏｕｓｅ」という言葉を含んだアンカーテキストに現れる場合、「ｈｏｕｓｅ」は、「ｇｒａｎｄｅ」にマッピングされるよりも、より適切に「ｃａｓａ」にマッピングされることを決定することは可能であり得る。例えば、既に、「ｂｉｇ」が非常に高確率で、および、十分に大きなデータのセットにわたって、「ｇｒａｎｄｅ」にマッピングされるように決定されている場合（および、アンカーテキストが同意語のリストで構成されてない場合）、次いで、ｈｏｕｓｅ－ｔｏ－ｃａｓａのマッピングは、たとえ、「ｈｏｕｓｅ」または「ｃａｓａ」を含むアンカーが要領を得ない場合であっても、ｈｏｕｓｅ－ｔｏ－ｇｒａｎｄｅのマッピングを介して、優先を与えられ得る。

【００７５】

変換の正確さおよび／または検索結果の有用性はまた、ユーザのクエリセッションの履歴を検証することによって改善され得る。例えば、多くの場合、システムは、ユーザが入力した以前のクエリを知っている（例えば、クッキーまたはサーバ上のユーザアカウントに格納された情報などを介して）。この履歴データは、そのユーザからのクエリの、可能な意味をランクするために使用され得、飛行に関連するものから、フィッシングに関するクエリに対して「ｂａｎｋ」を明確にする。したがって、このプロセスは、可能な変換のセットを狭めるために使用され得る。一部の実施形態において、ユーザインターフェースにて、「Ｘの検索を意図しますか？」（ここで、Ｘは所定の変換の優先を意味する）などのメッセージに関連して、それらを表示することによって、これらを示唆し得、その一方で、結果の第１のページにおいて、可能な再公式化のそれぞれからの結果の一部を表示する。ユーザが「…意図しますか？」によって示唆された代替のうちの一つを選択するか、結果ページに提示された結果のうちの一つを選択する場合、システムは、クエリの言葉の適当な変換、およびユーザの適当な検索バイアスに関する追加的な証拠を得る。これらの信号の両方は、次いで、システムによって利用され得、一般的な場合、およびユーザに特殊な場合の両方において、言葉のマッピングの適当な根拠を更新する（例えば、確率辞書において）。

【００７６】

Ｄ．結論

上記されたように、本発明と合致する方法およびシステムは、曖昧な検索結果に応じて検索結果を提供し、言葉を他の文字セットおよび／または言語に変換するために使用され得る。様々な変換および検索技術、ならびににシステムが記載されている。しかしながら、前述の記載は、例示のために提示されるものであり、多くの修正および変更が、上記の教示に照らし合わせて、または、本発明の実施を通して、可能であることは理解されたい。例えば、前述の記載はクライアント－サーバ構成に基づいているが、当業者は、ピアツーピア構成もまた、本発明に合致して使用され得ることは理解されたい。さらに、記載された実施はソフトウェアを含むが、本発明は、ハードウェアとソフトウェアとの組み合わせ、またはハードウェアのみとして実施され得る。付け加えて、本発明の局面は、メモリに格納されるように記載されたが、当業者は、これらの局面もまた、ハードディスク、フロッピーディスク（登録商標）、またはＣＤ－ＲＯＭなどのような二次的格納装置、インターネットからの搬送波、あるいは、ＲＡＭまたはＲＯＭの他の形式などの、他のタイプのコンピュータ可読媒体に格納され得る。本発明の範囲は、それゆえ、請求項およびそれらの均等物によって定義される。

【図面の簡単な説明】

【００７７】

【図１】本発明と合致する方法および装置がインプリメントされ得るシステムのブロック図を例示する。

【図２】本発明と合致する、クライアントデバイスのブロック図を例示する。

【図３】３つの文書を例示する図である。

【図４ａ】従来の英数字のインデックスを例示する。

【図４ｂ】従来の英数字の検索クエリに応じて、検索結果を提供するフロー図である。

【図５ａ】曖昧な検索クエリに応じた検索結果を提供するための、本発明に合致したフロー図を例示する。
【図５ｂ】数字の情報を数字の情報にマッピングする図を例示する。
【図５ｃ】（記載なし）
【図６】曖昧な検索クエリに応じた検索結果を提供するための、本発明に合致した別のフロー図を例示する。
【図７】本発明の実施形態に従い、検索を実行する方法を例示する。
【図８】文字セットの変換の確率辞書を例示する。
【図９】確率辞書を構築するためのパラレルアンカーテキストの使用を例示する。
【図１０】アンカーテキストを使用してリンクされた文書の集まりを例示する。
【図１１】図１１Ａおよび図１１Ｂは、図１０に示されたアンカーテキストに基づく適当な変換の計算を例示する。
【図１２】例示的な言葉の変換と関連した確率分配を示す。

【図１】

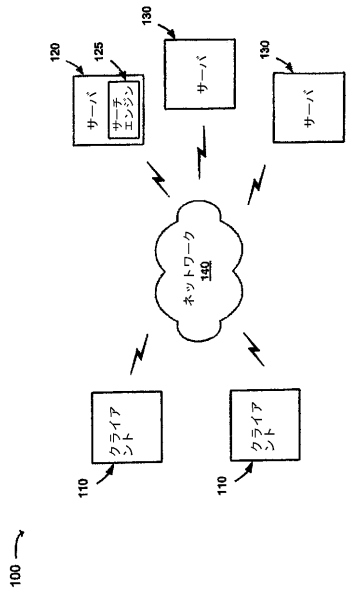


FIG. 1

【図２】

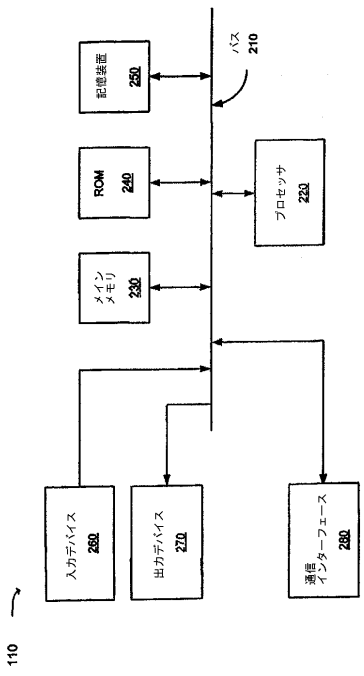


FIG. 2

【図３】

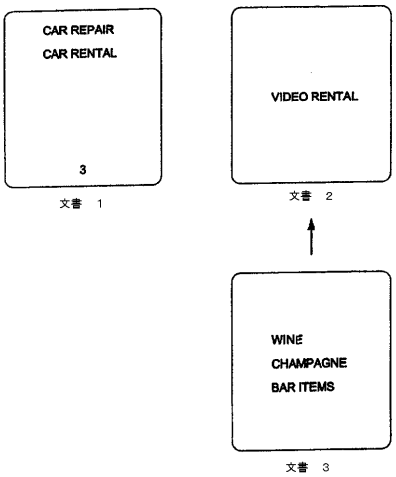


FIG. 3

【図４ａ】

言葉	位置（文書）
3	文書１
BAR	文書３
CAR	文書１
CHAMPAGNE	文書３
ITEMS	文書３
RENTAL	文書１および２
REPAIR	文書１
VIDEO	文書２
WINE	文書３

FIG. 4A

【図４ｂ】

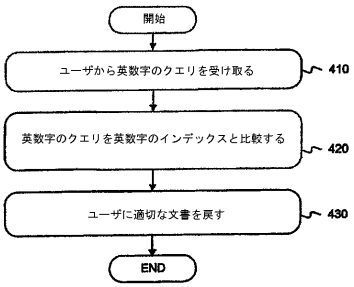


FIG. 4B

【図５ａ】

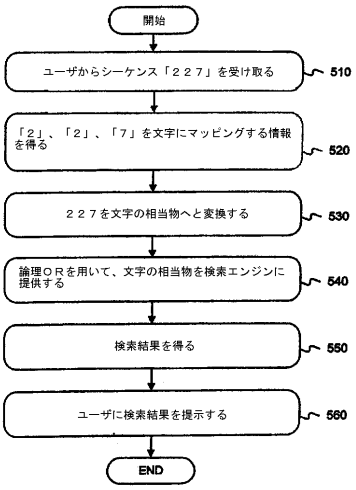


FIG. 5A

【図５Ｂ】

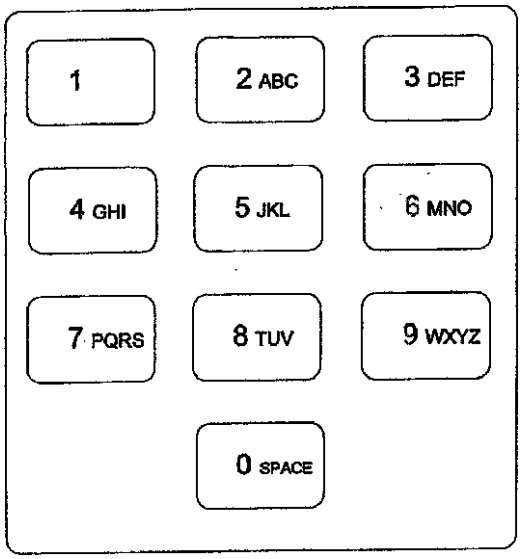


FIG. 5B

【図５ｃ】

言葉	位置（文書）
3	文書１
227	文書１および３
242672463	文書３
48367	文書３
736825	文書１および２
737247	文書１
84336	文書２
8463	文書３

FIG. 5C

【図６】

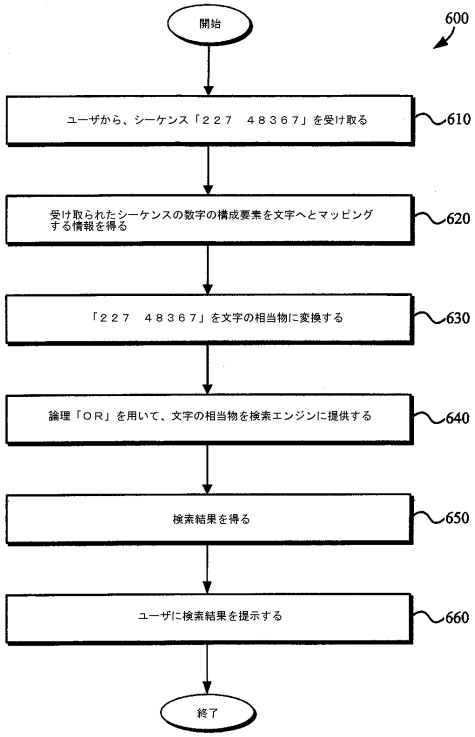


FIG. 6

【図７】

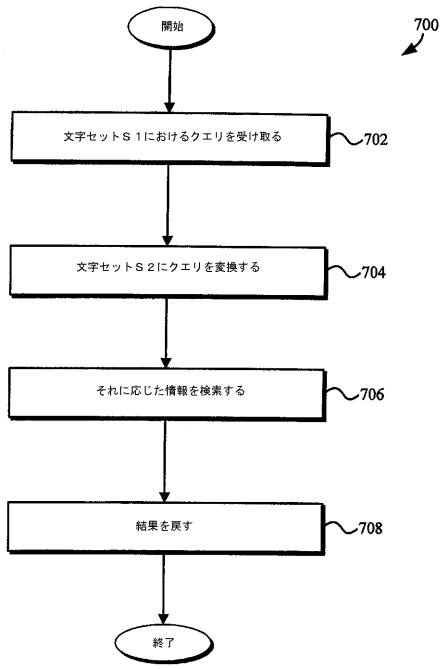
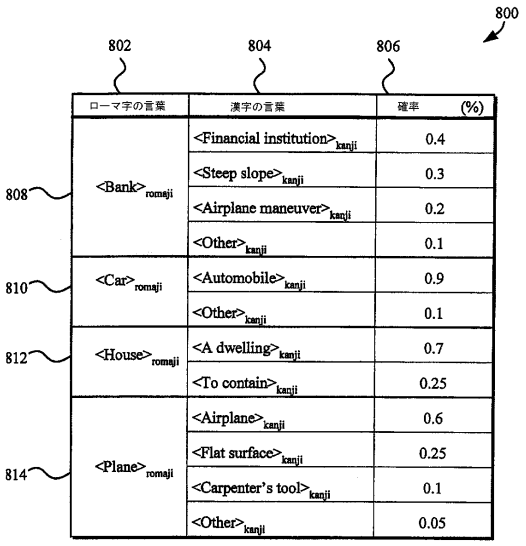


FIG. 7

【図８】



	ローマ字の言葉	漢字の言葉	確率 (%)
808	<Bank>romaji	<Financial institution>kanji	0.4
		<Steep slope>kanji	0.3
		<Airplane maneuver>kanji	0.2
		<Other>kanji	0.1
810	<Car>romaji	<Automobile>kanji	0.9
		<Other>kanji	0.1
812	<House>romaji	<A dwelling>kanji	0.7
		<To contain>kanji	0.25
814	<Plane>romaji	<Airplane>kanji	0.6
		<Flat surface>kanji	0.25
		<Carpenter's tool>kanji	0.1
		<Other>kanji	0.05

FIG. 8

【図９】

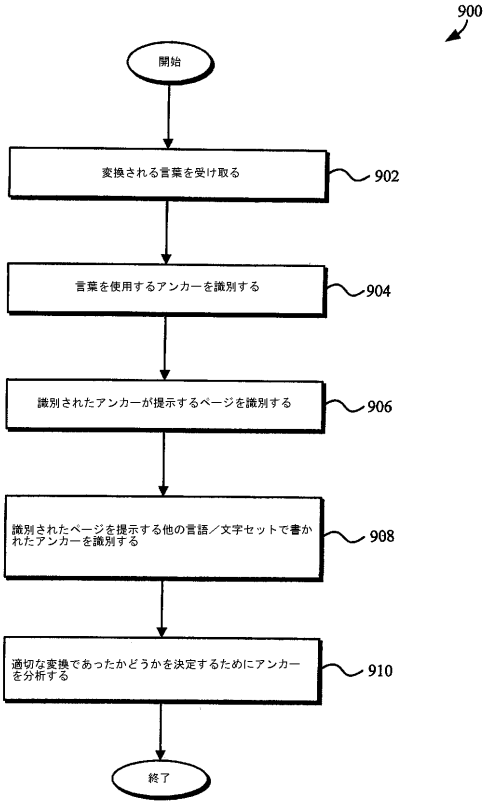


FIG. 9

【図１０】

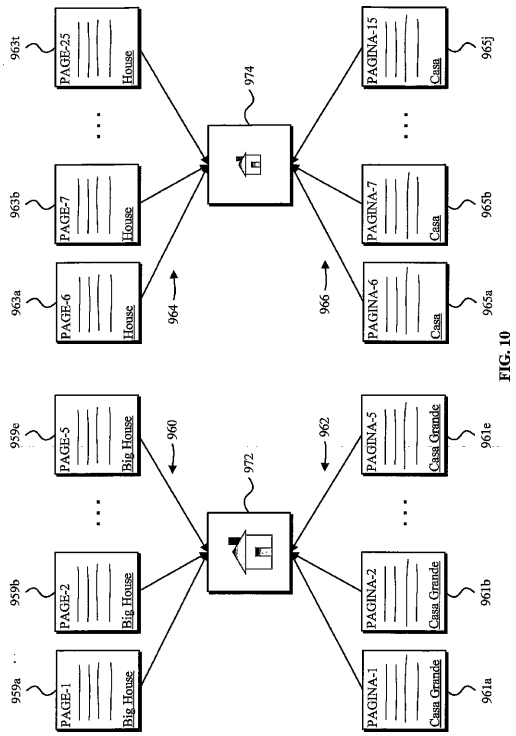
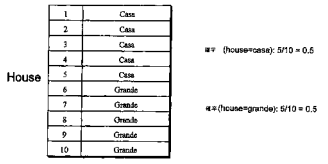


FIG. 10

【図１１】

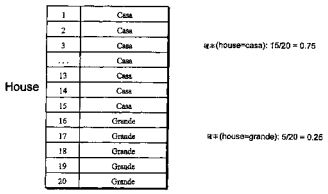


House	1 Casa
	2 Casa
	3 Casa
	4 Casa
	5 Casa
	6 Grande
	7 Grande
	8 Grande
	9 Grande
	10 Grande

(house=casa): 5/10 = 0.5

(house=grande): 5/10 = 0.5

FIG. 11A



House	1 Casa
	2 Casa
	3 Casa
	... Casa
	13 Casa
	14 Casa
	15 Casa
	16 Grande
	17 Grande
	18 Grande
	19 Grande
	20 Grande

(house=casa): 15/20 = 0.75

(house=grande): 5/20 = 0.25

FIG. 11B

【図１２】

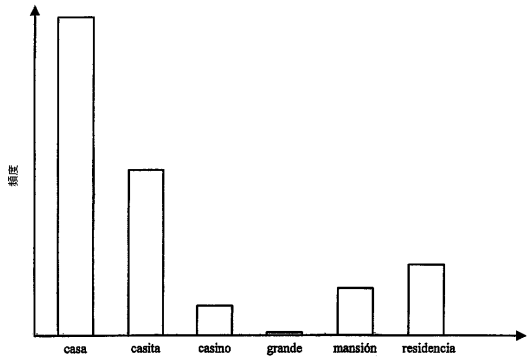


FIG. 12

フロントページの続き

(72)発明者 ポンテ, ジェイ エム.
アメリカ合衆国 カリフォルニア ９４０４３, マウンテン ビュー, マージョリー コート ２４３９
(72)発明者 サハミ, メヘラーン
アメリカ合衆国 カリフォルニア ９４０６３, レッドウッド シティ, フーバー ストリート ３２３８
(72)発明者 ゲマワット, サンジャイ
アメリカ合衆国 カリフォルニア ９４０４３, マウンテン ビュー, ノース レンフストルフ アベニュー １１１, ナンバー１８４
(72)発明者 バウアー, ジョン エー.
アメリカ合衆国 カリフォルニア ９４０４０, マウンテン ビュー, デル メディオ アベニュー ４１５, ナンバー８

審査官 波内 みさ

(56)参考文献 特開２０００－１６３４４１（ＪＰ，Ａ）
特開２００２－０２４２６６（ＪＰ，Ａ）
国際公開第２００３／０５８３７４（WO，Ａ１）
特開２００２－２２２１８９（ＪＰ，Ａ）
特開２００２－２５９３７４（ＪＰ，Ａ）
菊井 玄一郎，言語の壁を越えて文書を検索する－クロスランゲージ情報検索－，人工知能学会誌 第１５巻 第４号，日本，社団法人人工知能学会，２０００年 ７月 １日，第15巻 第4号，550～558
乾 健太郎，言語表現を言い換える技術，言語処理学会第８回年次大会チュートリアル資料，日本，言語処理学会，２００２年 3月１７日，1～21

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30
G06F 17/28