(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2009-528636
(P2009-528636A)
(43)公表日　平成21年8月6日(2009.8.6)

| (51)Int.Cl. | | F I | | | テーマコード（参考） |
|---|---|---|---|---|---|
| G06F 17/30 | (2006.01) | G06F | 17/30 | 330B | 5B075 |
| | | G06F | 17/30 | 320C | |

審査請求 有　予備審査請求 未請求　（全 31 頁）

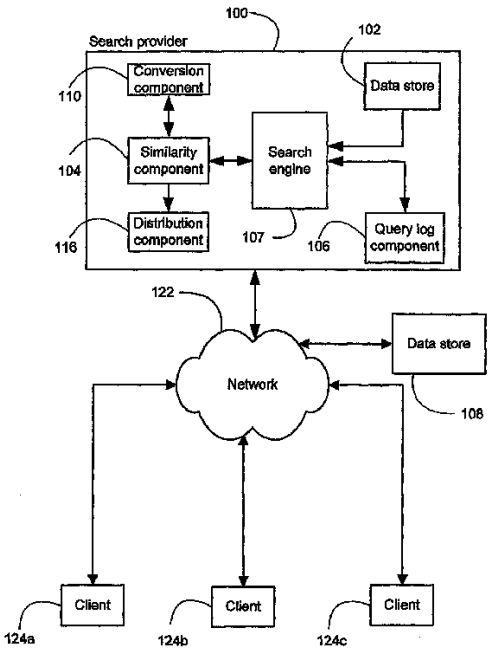
| | |
|---|---|
| (21)出願番号 | 特願2008-557464 (P2008-557464) |
| (86)(22)出願日 | 平成19年2月27日 (2007.2.27) |
| (85)翻訳文提出日 | 平成20年10月27日 (2008.10.27) |
| (86)国際出願番号 | PCT/US2007/062876 |
| (87)国際公開番号 | WO2007/101194 |
| (87)国際公開日 | 平成19年9月7日 (2007.9.7) |
| (31)優先権主張番号 | 11/365,315 |
| (32)優先日 | 平成18年2月28日 (2006.2.28) |
| (33)優先権主張国 | 米国 (US) |

(71)出願人　501438485
ヤフー！　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　９４０８９　サニーヴェイル　ファースト　アヴェニュー　７０１
(74)代理人　100082005
弁理士　熊倉　禎男
(74)代理人　100067013
弁理士　大塚　文昭
(74)代理人　100086771
弁理士　西島　孝喜
(74)代理人　100109070
弁理士　須田　洋之
(74)代理人　100120525
弁理士　近藤　直樹

最終頁に続く

(54)【発明の名称】複数の書記体系を有する言語に対する関連のクエリーを識別するためのシステム及び方法

(57)【要約】
【課題】　複数の書記体系を有する言語に従って書かれた所定の検索クエリーに関連する１つ又はそれよりも多くのクエリーを識別する方法及びシステムを提供する。
【解決手段】　所定のクエリーに関連する１つ又はそれよりも多くのクエリーを識別するためのシステム及び方法。本発明の方法は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーを受け取る段階を含む。複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーの候補セットが識別される。受け取られたクエリーに対する１つ又はそれよりも多くのクエリーの類似性を示すスコアが、候補セット内の１つ又はそれよりも多くのクエリーに対して計算される。
【選択図】　図１



【特許請求の範囲】
【請求項１】
所定のクエリーに関連する１つ又はそれよりも多くのクエリーを識別する方法であって、
複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーを受け取る段階と、
複数の書記体系を有する前記言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーの候補セットを識別する段階と、
前記候補セット内の前記１つ又はそれよりも多くのクエリーに対して、前記受け取られたクエリーに対する該１つ又はそれよりも多くのクエリーの類似性を示すスコアを計算する段階と、
を含むことを特徴とする方法。
【請求項２】
前記クエリーを受け取る段階は、１つ又はそれよりも多くの日本語書記体系の組合せに従って書かれたクエリーを受け取る段階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】
前記クエリーの候補セットを識別する段階は、前記受け取られたクエリーに関連する１組の１つ又はそれよりも多くのクエリーを識別する段階を含むことを特徴とする請求項１に記載の方法。
【請求項４】
前記受け取られたクエリーに関連する前記クエリーの候補セットを識別する段階は、１つ又はそれよりも多くのクエリーログにおいて示されるような該受け取られたクエリーに関連する１つ又はそれよりも多くのクエリーを識別する段階を含むことを特徴とする請求項３に記載の方法。
【請求項５】
前記クエリーを受け取る段階は、日本語ひらがな書記体系に従って書かれたクエリーを受け取る段階を含むことを特徴とする請求項１に記載の方法。
【請求項６】
前記クエリーを受け取る段階は、日本語カタカナ書記体系に従って書かれたクエリーを受け取る段階を含むことを特徴とする請求項１に記載の方法。
【請求項７】
前記クエリーを受け取る段階は、日本語かな書記体系に従って書かれたクエリーを受け取る段階を含むことを特徴とする請求項１に記載の方法。
【請求項８】
前記クエリーを受け取る段階は、日本語ローマ字書記体系に従って書かれたクエリーを受け取る段階を含むことを特徴とする請求項１に記載の方法。
【請求項９】
前記クエリーを受け取る段階は、日本語ＪＡＳＣＩＩ書記体系に従って書かれたクエリーを受け取る段階を含むことを特徴とする請求項１に記載の方法。
【請求項１０】
前記クエリーを受け取る段階は、日本語漢字書記体系に従って書かれたクエリーを受け取る段階を含むことを特徴とする請求項１に記載の方法。
【請求項１１】
前記クエリーを受け取る段階は、語句を含む１組の用語を受け取る段階を含むことを特徴とする請求項１に記載の方法。
【請求項１２】
前記候補セット内の前記１つ又はそれよりも多くのクエリーに対するスコアを計算する段階は、前記受け取られたクエリーに対する前記候補からの所定のクエリーの意味における類似性を示すスコアを計算する段階を含むことを特徴とする請求項１に記載の方法。
【請求項１３】

スコアを計算する段階は、
前記受け取られたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記候補セットから選択されたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記受け取られたクエリーと前記候補セットからの前記選択されたクエリーの間の文字編集距離を計算する段階と、
を含む、
ことを特徴とする請求項１に記載の方法。
【請求項１４】
スコアを計算する段階は、
前記受け取られたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記候補セットから選択されたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記受け取られたクエリーと前記候補セットからの前記選択されたクエリーからスペース文字を取り除く段階と、
前記受け取られたクエリーと前記候補セットからの前記選択されたクエリーとの間の文字編集距離を計算する段階と、
を含む、
ことを特徴とする請求項１に記載の方法。
【請求項１５】
スコアを計算する段階は、
前記受け取られたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記候補セットから選択されたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記受け取られたクエリー及び前記選択されたクエリー内のスペースで区切られた固有の共起語の数を識別する段階と、
前記受け取られたクエリーと前記選択されたクエリーの両方におけるスペースで区切られた固有の語の総数を識別する段階と、
両方のクエリー内のスペースで区切られた固有の共起語の前記数とスペースで区切られた固有の語の前記総数との商を計算する段階と、
数値１と前記計算された商との間の差を計算する段階と、
を含む、
ことを特徴とする請求項１に記載の方法。
【請求項１６】
スコアを計算する段階は、数字が、前記受け取られたクエリー又は前記候補セットから選択されたクエリーに固有であるか否かを識別する段階を含むことを特徴とする請求項１に記載の方法。
【請求項１７】
スコアを計算する段階は、
前記受け取られたクエリー及び前記候補セットから選択されたクエリー内の共起する日本語漢字文字の数を識別する段階と、
前記受け取られたクエリー及び前記候補セットからの前記選択されたクエリー内の固有の日本語漢字文字の総数を識別する段階と、
共起する日本語漢字文字の前記数と固有の日本語漢字文字の前記総数との商を計算する段階と、
数値１と前記計算された商との間の差を計算する段階と、
を含む、

ことを特徴とする請求項１に記載の方法。
【請求項１８】
スコアを計算する段階は、
前記受け取られたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記候補セットから選択されたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、
前記受け取られたクエリーと前記選択されたクエリーとが共通して有するローマ字の数を計算する段階と、
を含む、
ことを特徴とする請求項１に記載の方法。
【請求項１９】
スコアを計算する段階は、前記受け取られたクエリー又は前記候補セットから選択されたクエリーのいずれかが、非ローマ字文字を包含するか否かを識別する段階を含むことを特徴とする請求項１に記載の方法。
【請求項２０】
スコアを計算する段階は、
前記受け取られたクエリーの１つ又はそれよりも多くの日本語漢字文字を日本語かな文字に変換する段階と、
前記候補セットから選択されたクエリーの１つ又はそれよりも多くの日本語漢字文字を日本語かな文字に変換する段階と、
前記受け取られたクエリー及び前記候補セットからの前記選択されたクエリーから全ての非日本語文字を取り除く段階と、
前記受け取られたクエリーと前記候補セットからの前記選択されたクエリーとの間の文字編集距離を計算する段階と、
を含む、
ことを特徴とする請求項１に記載の方法。
【請求項２１】
スコアを計算する段階は、前記候補セットからの選択されたクエリーが、１つ又はそれよりも多くのクエリーログ内で前記受け取られたクエリーに続く頻度と、該１つ又はそれよりも多くのクエリーログ内の該受け取られたクエリーの頻度との商を計算する段階を含むことを特徴とする請求項１に記載の方法。
【請求項２２】
分配のために前記候補セットからの前記クエリーの１つ又はそれよりも多くを選択する段階を含むことを特徴とする請求項１に記載の方法。
【請求項２３】
分配のために前記候補セットからの前記クエリーの１つ又はそれよりも多くを選択する段階は、所定の閾値を超えるスコアを有する１つ又はそれよりも多くのクエリーを選択する段階を含むことを特徴とする請求項２２に記載の方法。
【請求項２４】
所定の閾値を超えるスコアを有する前記候補セットからの前記１つ又はそれよりも多くのクエリーを分配する段階を含むことを特徴とする請求項１に記載の方法。
【請求項２５】
前記候補セットからの前記１つ又はそれよりも多くのクエリーを分配する段階は、該１つ又はそれよりも多くのクエリーをウェブページに組み込む段階を含むことを特徴とする請求項２４に記載の方法。
【請求項２６】
所定のクエリーに関連する１つ又はそれよりも多くのクエリーを識別するためのシステムであって、
複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたク

エリーを受け取り、かつ
複数の書記体系を有する前記言語の１つ又はそれよりも多くの書記体系に従って書かれた１つ又はそれよりも多くのクエリーの候補セットを識別する、
ように作動する検索エンジンと、
前記受け取られたクエリーと前記候補セット内の前記１つ又はそれよりも多くのクエリーとを１つ又はそれよりも多くの文書フォーマットに変換するように作動する変換構成要素と、
前記受け取られたクエリーに対する前記１つ又はそれよりも多くのクエリーの類似性を示す、前記候補セット内の前記１つ又はそれよりも多くのクエリーに対するスコアを計算するように作動する類似性構成要素と、
を含むことを特徴とするシステム。
【請求項２７】
前記検索エンジンは、１つ又はそれよりも多くの日本語書記体系に従って書かれたクエリーを受け取るように作動することを特徴とする請求項２６に記載のシステム。
【請求項２８】
前記検索エンジンは、前記受け取られたクエリーに関連する１つ又はそれよりも多くのクエリーから成る候補セットを識別するように作動することを特徴とする請求項２６に記載のシステム。
【請求項２９】
前記検索エンジンは、前記受け取られたクエリーに関連する１つ又はそれよりも多くのクエリーを識別するために１つ又はそれよりも多くのクエリーログを検索するように作動することを特徴とする請求項２８に記載のシステム。
【請求項３０】
前記変換構成要素は、１つ又はそれよりも多くの書記体系に従ってクエリーを１つ又はそれよりも多くの文書フォーマットに変換するように作動することを特徴とする請求項２６に記載のシステム。
【請求項３１】
前記類似性構成要素は、前記受け取られたクエリーに対する前記候補セットから選択されたクエリーの意味における類似性を示すスコアを計算するように作動することを特徴とする請求項２６に記載のシステム。
【請求項３２】
前記類似性構成要素は、前記受け取られたクエリーと前記候補セットから選択されたクエリーとの間の文字編集距離を計算するように作動することを特徴とする請求項２６に記載のシステム。
【請求項３３】
前記類似性構成要素は、
前記受け取られたクエリー及び前記選択されたクエリー内のスペースで区切られた固有の共起語の数を識別し、
前記受け取られたクエリー及び前記選択されたクエリーの両方におけるスペースで区切られた固有の語の総数を識別し、
両方のクエリー内のスペースで区切られた固有の共起語の前記数とスペースで区切られた固有の語の前記総数との商を計算し、かつ
数値１と前記計算された商との間の差を計算する、
ように作動する、
ことを特徴とする請求項２６に記載のシステム。
【請求項３４】
前記類似性構成要素は、数字が、前記受け取られたクエリー又は前記候補セットから選択されたクエリーに固有であるか否かを識別するように作動することを特徴とする請求項２６に記載のシステム。
【請求項３５】

前記類似性構成要素は、
前記受け取られたクエリー及び前記候補セットから選択されたクエリー内の共起する日本語漢字文字の数を識別し、
前記受け取られたクエリー及び前記候補セットからの前記選択されたクエリー内の固有の日本語漢字文字の総数を識別し、
共起する日本語漢字文字の前記数と固有の日本語漢字文字の前記総数との商を計算し、
数値１と前記計算された商との間の差を計算する、
ように作動する、
ことを特徴とする請求項２６に記載のシステム。
【請求項３６】
前記類似性構成要素は、前記受け取られたクエリーと前記候補セットから選択されたクエリーとが共通して有する文字の数を計算するように作動することを特徴とする請求項２６に記載のシステム。
【請求項３７】
前記類似性構成要素は、前記受け取られたクエリー又は前記候補セットから選択されたクエリーが、所定の書記体系の１つ又はそれよりも多くの文字を包含するか否かを識別するように作動することを特徴とする請求項２６に記載のシステム。
【請求項３８】
前記類似性構成要素は、前記候補セットから選択されたクエリーが、１つ又はそれよりも多くのクエリーログ内で前記受け取られたクエリーに続く頻度と、該１つ又はそれよりも多くのクエリーログ内の該受け取られたクエリーの頻度との商を計算するように作動することを特徴とする請求項２６に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
著作権通知
本特許文書の開示の部分は、著作権保護された材料を包含する。著作権所有者は、「特許及び商標事務所」特許ファイル又は記録に現れる場合の本特許文書又は特許開示の他者によるファクシミリ複製に異議はないが、それ以外は全ての著作権を保有するものである。
【０００２】
関連出願への相互参照
本出願は、各々が本明細書においてその全内容が引用により組み込まれている以下の係属中の出願に関連する。
・２００５年８月１０日出願の「代替検索クエリーを判断するためのシステム及び方法」という名称の米国特許出願出願番号第１１／２００、８５１号、及び
・２００５年１１月９日出願の「モジュラー最適化動的セット」という名称の米国特許仮出願第６０／７３６、１３３号。
【０００３】
本発明は、一般的に、複数の書記体系を有する言語に従って書かれた所定の検索クエリーに関連する１つ又はそれよりも多くのクエリーを識別する方法及びシステムを提供する。より具体的には、本発明は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系の組合せに従って書かれた検索クエリーを受け取り、クエリーの候補セットから１つ又はそれよりも多くの関連するクエリーを識別する方法及びシステムを提供する。
【背景技術】
【０００４】
「ワールド・ワイド・ウェブ」（ウェブ）を通じてユーザに利用可能な「インターネット」及び多数のウェブページ、メディアコンテンツ、広告などの進歩と共に、ウェブから該当する情報を取得するための能率化された手法をユーザに供給する必要が生じている。このような情報を取得するユーザの必要性を満たすために、検索システム及び処理が開発

されている。このような技術の例は、Ｙａｈｏｏ！、Ｇｏｏｇｌｅ、及び他の検索プロバイダウェブサイトを通じてアクセス可能である。

【０００５】

現在、ユーザは、コンテンツを検索して取り出すためにワイドエリアネットワーク、例えば「インターネット」へのアクセスを備えたクライアントデバイス（パーソナルコンピュータ（ＰＣ）、ＰＤＡ、スマートフォンなど）を利用することができる。一般的に、ユーザは、クライアントデバイスを通じてクエリーを入力し、検索処理は、クエリーに関連したリンク、文書、ウェブページ、広告などのような１つ又はそれよりも多くのコンテンツの項目を戻す。所定のクエリーに応答して戻されるコンテンツの項目は、ユーザが実際に求めていたサブジェクト又はトピックに密接に関連することもあり、又は全く関連しないこともある。取り出されたコンテンツの項目が所定のクエリーにどのくらい近く関連するかに基づいて測ることができる所定の検索の成功は、検索クエリーの適正な解釈に大きく依存する場合がある。

【０００６】

クエリーは、１つ又はそれよりも多くの語及び語句から作られる。しかし、人間ユーザによって入力されたクエリーは、所定のユーザが求めているコンテンツを適切に表わせないことが多い。更に、ユーザは、求めているコンテンツの一般的な又は漠然とした知識しか持たない可能性がある。例えば、ユーザが、テレビで宣伝された製品に対してＹａｈｏｏ！検索エンジンを使用して検索を行いたい場合がある。ユーザは、製品の名前、製造業者などを知らない場合があり、製品を一般的に表現することができるのみである場合がある。従って、ユーザによって作成されたクエリーが広義すぎて、ユーザによって求められたコンテンツに全く関係ないコンテンツ項目の検索をもたらす。同様に、ユーザによって選択されたクエリー用語は、製品を適切に表現できない場合があり、たとえあったとしてもごく少ないコンテンツ項目の取り出しをもたらす。

【０００７】

所定のクエリーに関連すると考えられるクエリーの候補セットを生成する現在の技術は公知である。例えば、ユーザは、「アップル（登録商標）ＭＰ３プレーヤ」というクエリーを入力することができ、「ＩＰＯＤ（登録商標）」、「Ｉｔｕｎｅｓ（登録商標）」などのような１つ又はそれよりも多くの関連するクエリーを表示される。しかし、検索プロバイダは、所定のクエリーに意味において最も該当するか又は密接に関連する１つ又はそれよりも多くのクエリーをクエリーの候補セットから識別するという問題を呈示される。更に、日本語のようなある一定の言語は、複数の書記体系を有し、これは、所定のクエリーに意味において最も該当するか又は類似のクエリーをクエリーの候補セットから識別するという複雑さを更に増大させる。例えば、検索エンジンに提出された単一の日本語のクエリーは、漢字、カタカナ、ひらがな、ＪＡＳＣＩＩ、ＡＳＣＩＩなどのような１つ又はそれよりも多くの日本語書記体系の様々な組合せに従って書かれる場合がある。日本語の漢字書記体系に従って書かれたクエリーは、日本語のカタカナ及びひらがな書記体系に従って書かれたクエリーとは全く異なるように見えるであろうが、２つのクエリーは、非常に類似又は同一の意味を有する場合がある。

【０００８】

更に、Ｙａｈｏｏ！、ＭＳＮ、又はＧｏｏｇｌｅのような検索プロバイダは、広告主が、クエリーに応答して１つ又はそれよりも多くの広告を表示させるために用語に対して入札することができる入札市場を利用することができる。例えば、１つ又はそれよりも多くの広告主は、ラップトップコンピュータに対する１つ又はそれよりも多くの広告を表示したい場合があり、従って、「ノートブックコンピュータ」という用語に対して入札することができる。しかし、「ノートブックコンピュータ」という用語は、日本語のような複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書くことができる。例えば、「ノートブックコンピュータ」という用語は、日本語ひらがな書記体系、日本語カタカナ書記体系などに従って書くことができる。

【０００９】

ユーザは、日本語カタカナ書記体系に従って書かれた「ノートブックコンピュータ」という用語を含むクエリーをＹａｈｏｏ！のような所定の検索プロバイダに提出することができる。カタカナ用語「ノートブックコンピュータ」に対する関連の入札値を有する１つ又はそれよりも多くの広告が取り出されて、ユーザに表示することができる。入札市場では、カタカナ語「ノートブックコンピュータ」に対して最も大きな入札値を供給した広告主に関連付けられた広告が、ウェブページの最も目立つ、例えば、広告のランク付けリストで一番にランク付けされ、所定の検索結果ページの最上部に表示される位置に表示される、等々である。

【００１０】

ユーザが、表示された広告の１つ又はそれよりも多くを選択した場合、検索プロバイダは、選択された広告に関連付けられた広告主に広告主の指し値に基づく金額を請求することなどにより、ユーザの選択を貨幣化することができる。しかし、１つ又はそれよりも多くの用語に対する関連の指し値を有する広告のみを取り出して表示することは、所定の検索プロバイダへの収入のかなりの損失をもたらすであろう。例えば、ユーザが、１つ又はそれよりも多くの広告主によって入札されていない用語から成るクエリーを入力した場合、検索プロバイダは、ユーザにどの広告も戻すことができず、ユーザがどの結果も選択できないことになるので、検索プロバイダに収入の損失をもたらす。上述の例に関して、ユーザによって入力されたクエリーがカタカナ用語「ノートブックコンピュータ」を含まず、しかし、代わりにひらがな用語「らっぷとっぷこんぴゅーた」を含んでいた場合、検索プロバイダは、カタカナクエリー「ラップトップコンピュータ」とひらがなクエリー「のーとぶっくこんぴゅーた」の意味の類似性にも関わらず、目標の広告を適正に表示できない場合がある。

【００１１】

所定のクエリーに意味において同一又は類似の１つ又はそれよりも多くのクエリーをクエリーの候補セットから識別するための技術は存在するが、既存の技術は、単一の書記体系に従って書かれた言語に制限される。従って、現在の技術は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたオリジナルのクエリーに意味において最も該当するか又は密接に関連するクエリーの識別を提供できない。既存の技術に関連した欠点を克服するために、本発明は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた所定の検索クエリーに関して意味において最も類似する１つ又はそれよりも多くのクエリーを関連するクエリーの候補セットから識別するためのシステム及び方法を提供する。

【００１２】

【特許文献１】米国特許出願出願番号第１１／２００、８５１号

【特許文献２】米国特許仮出願第６０／７３６、１３３号

【発明の開示】

【００１３】

本発明は、所定のクエリーに関連する１つ又はそれよりも多くのクエリーを識別する方法及びシステムに関連する。本発明の方法は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーを受け取る段階を含む。本発明の一実施形態によると、受け取られたクエリーは、日本語のひらがな、カタカナ、かな、ローマ字、ＪＡＳＣＩＩ、及び漢字書記体系を含む１つ又はそれよりも多くの日本語書記体系の組合せに従って書かれたクエリーを含む。

【００１４】

受け取られたクエリーに付随する複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーの候補セットが識別される。本発明の一実施形態によると、クエリーの候補セットは、１つ又はそれよりも多くのクエリーログにおいて指示されたような受け取られたクエリーに関連する１つ又はそれよりも多くのクエリーを含む。

本方法は、受け取られたクエリーに対する１つ又はそれよりも多くのクエリーの類似性

を示す候補セット内の１つ又はそれよりも多くのクエリーに対するスコアを計算する段階を更に含む。候補セット内の１つ又はそれよりも多くのクエリーに対して計算されたスコアは、受け取られたクエリーに対する候補セットからの所定のクエリーの意味における類似性を示している。本発明の一実施形態によると、スコアを計算する段階は、各クエリーの１つ又はそれよりも多くの文字をローマ字に変換した後で、受け取られたクエリーと候補セットから選択されたクエリーとの間の文字編集距離を計算する段階を含む。本発明の別の実施形態によると、スコアを計算する段階は、各クエリーの１つ又はそれよりも多くの文字をローマ字に変換して各クエリーからスペース文字を取り除いた後で、受け取られたクエリーと候補セットから選択されたクエリーとの間の文字編集距離を計算する段階を含む。本発明の更に別の実施形態によると、スコアを計算する段階は、受け取られたクエリーと候補セットから選択されたクエリーとの文字をローマ字に変換する段階、及び１と、受け取られたクエリーと選択されたクエリーにおけるスペースで区切られた固有の共起語の数と両方のクエリーにおけるスペースで区切られた固有の語の総数との商との間の差を計算する段階を含む。

【００１５】

本発明の更に別の実施形態によると、スコアを計算する段階は、数字が、受け取られたクエリーと候補セットから選択されたクエリーとに固有のものであるか否かを識別する段階を含む。更に別の実施形態によると、スコアを計算する段階は、値１と、受け取られたクエリーと候補セットからの選択されたクエリーとにおける共起日本語漢字文字の数と、受け取られたクエリーと候補セットからの選択されたクエリーとにおける固有の日本語漢字文字の総数との商との間の差を計算する段階を含む。本発明の別の実施形態によると、スコアを計算する段階は、受け取られたクエリー及び候補セットから選択されたクエリーの１つ又はそれよりも多くの文字をローマ字に変換する段階と、これらのクエリーが共通して有するローマ字の数を計算する段階とを含む。本発明の更に別の実施形態によると、スコアを計算する段階は、受け取られたクエリー又は候補セットからの選択されたクエリーのいずれかが非ローマ字文字を包含するか否かを識別する段階を含む。本発明の更に別の実施形態によると、スコアを計算する段階は、各クエリーの日本語漢字文字を日本語かな文字に変換して各クエリーから全ての非日本語文字を取り除いた後で、受け取られたクエリーと候補セットからの選択されたクエリーとの間の文字編集距離を計算する段階を含む。更に別の実施形態によると、スコアを計算する段階は、候補セットからの選択されたクエリーが１つ又はそれよりも多くのクエリーログ内で受け取られたクエリーに続く頻度と、１つ又はそれよりも多くのクエリーログ内の受け取られたクエリーの頻度との商を計算する段階を含む。

【００１６】

本方法は、分配のために候補セットからクエリーの１つ又はそれよりも多くを選択する段階を更に含む。本発明の一実施形態によると、分配のために候補セットから選択された１つ又はそれよりも多くのクエリーは、所定の閾値を超えるスコアを有するクエリーを含む。分配のために選択された１つ又はそれよりも多くのクエリーは、分配することができる。本発明の一実施形態によると、分配のために選択されたクエリーは、１つ又はそれよりも多くのウェブページに組み込まれる。

【００１７】

本発明は、所定のクエリーに関連する１つ又はそれよりも多くのクエリーを識別するためのシステムにも関連する。本発明のシステムは、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーを受け取るように作動する検索エンジンを含む。本発明の一実施形態によると、検索エンジンは、１つ又はそれよりも多くの日本語書記体系に従って書かれたクエリーを受け取るように作動する。検索エンジンは、受け取られたクエリーに付随する複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた１つ又はそれよりも多くのクエリーの候補セットを識別するように更に作動する。本発明の一実施形態によると、検索エンジンは、１つ又はそれよりも多くのクエリーログにおいて指示されたような受け取られたクエリーに関連する１つ

又はそれよりも多くのクエリーから成る候補セットを識別するように作動する。
変換構成要素は、受け取られたクエリーと候補セット内の１つ又はそれよりも多くのクエリーとを１つ又はそれよりも多くの文書フォーマットに変換するように作動する。本発明の一実施形態によると、変換構成要素は、１つ又はそれよりも多くの書記体系に従ってクエリーを１つ又はそれよりも多くの文書フォーマットに変換するように作動する。
【００１８】
類似性構成要素は、受け取られたクエリーに対する１つ又はそれよりも多くのクエリーの類似性を示すスコアを候補セット内の１つ又はそれよりも多くのクエリーに対して計算するように作動する。類似性構成要素は、受け取られたクエリーに対して候補セットからの選択されたクエリーの意味における類似性を示すスコアを計算するように作動する。本発明の一実施形態によると、類似性構成要素は、受け取られたクエリーと候補セットからの選択されたクエリーとの間の文字編集距離を計算するように作動する。本発明の更に別の実施形態によると、類似性構成要素は、１と、受け取られたクエリーと候補セットから選択されたクエリーとにおけるスペースで区切られた固有の共起語の数と両方のクエリーにおけるスペースで区切られた固有の語の総数との商との間の差を計算するように作動する。本発明の更に別の実施形態によると、類似性構成要素は、数字が、受け取られたクエリー又は候補セットからの選択されたクエリーに固有であるか否かを識別するように作動する。
【００１９】
別の実施形態によると、類似性構成要素は、１と、受け取られたクエリーと候補セットから選択されたクエリーとにおける共起日本語漢字文字の数と両方のクエリーにおける固有の日本語漢字文字の総数との商との間の差を計算するように作動する。本発明の更に別の実施形態によると、類似性構成要素は、受け取られたクエリーと候補セットからの選択されたクエリーとが共通して有する文字の数を計算するように作動する。本発明の更に別の実施形態によると、類似性構成要素は、受け取られたクエリー又は候補セットからの選択されたクエリーが、所定の書記体系の１つ又はそれよりも多くの文字を包含するか否かを識別するように作動する。更に別の実施形態によると、類似性構成要素は、候補セットからの選択されたクエリーが１つ又はそれよりも多くのクエリーログ内の受け取られたクエリーに続く頻度と、クエリーログ内の受け取られたクエリーの頻度との商を計算するように作動する。
【００２０】
本発明は、同じ参照が同じか又は対応する部分を示すものとする添付図面において例示的であって制限を意図しない図に例証される。
【発明を実施するための最良の形態】
【００２１】
以下の説明では、説明の一部を形成する添付の図面を参照し、図面には、本発明を実施することができる特定的な実施形態を例証によって示している。他の実施形態を利用することができること、及び本発明の範囲から逸脱することなく構造的な変更を行い得ることは理解されるものとする。
【００２２】
図１は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた所定のクエリーに関連する１つ又はそれよりも多くのクエリーを識別するためのシステムの一実施形態を示すブロック図である。図１の実施形態によると、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃは、「インターネット」のような１つ又はそれよりも多くのローカル及び／又はワイドエリアネットワークへの接続を含むことができるネットワーク１２２に通信することができるように連結されている。本発明の一実施形態によると、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃは、プロセッサ、一時及び永続記憶デバイス、入力／出力サブシステム、及び汎用パーソナルコンピュータを含む構成要素間の通信経路を提供するバスを含む汎用パーソナルコンピュータである。例えば、５１２ＭＢのＲＡＭ、４０ＧＢのハードドライブ記憶スペース、及びネットワーク

への「イーサネット（登録商標）」インタフェースを備えた３．５ＧＨｚ「Ｐｅｎｔｉｕｍ（登録商標）　４」パーソナルコンピュータである。

【００２３】

ネットワーク１２２に通信することができるように連結されたクライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃのユーザは、１つ又はそれよりも多くの用語を含む検索クエリーを検索プロバイダ１００に提出することができる。ユーザによってネットワーク１２２を通じて検索プロバイダ１００に提出された検索クエリーは、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた１つ又はそれよりも多くの文字、用語、又は語句を含むことができる。例えば、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃのユーザは、日本語漢字文字、日本語カタカナ文字、及びＪＡＳＣＩＩ文字を含むクエリーを作成することができる。同様に、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃのユーザは、日本語ローマ字文字、日本語ひらがな文字、及び数字を含むクエリーを作成することができる。例えば、ユーザは、日本語のカタカナ、ひらがな、漢字、及びＡＳＣＩＩ書記体系の組合せに従って書かれた以下のクエリー、すなわち、「１リットルの涙　沢尻エリカ」を提出することができる。

【００２４】

クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃのユーザによって提出された複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた文字及び用語を含むことができる１つ又はそれよりも多くの検索クエリーは、関連するクエリーの候補セットを識別するために検索プロバイダ１００で検索エンジン１０７によって使用される。関連するクエリーの候補セットを含む１つ又はそれよりも多くのクエリーは、所定のクエリーに関連する１つ又はそれよりも多くのクエリーを保持するようにそれぞれが作動する１つ又はそれよりも多くのローカル又はリモートデータ記憶装置１０２及び１０８に保持することができる。本発明の一実施形態によると、データ記憶装置１０２及び１０８は、１つ又はそれよりも多くのクエリー又は用語に関連するクエリーのセットを識別するエントリを備えた索引を保持するように作動する。データ記憶装置１０２及び１０８によって保持される索引は、関連する用語又はクエリーを示す人間が編集する情報で補足される。例えば、データ記憶装置１０２及び１０８内の索引エントリは、日本語のカタカナ、ひらがな、漢字、及びＡＳＣＩＩ書記体系に従って書かれたクエリー「１リットルの涙　沢尻エリカ」と、１つ又はそれよりも多くの日本語書記体系に従って書かれた１つ又はそれよりも多くの関連するクエリー又は用語とを含むことができる。

【００２５】

データ記憶装置１０２及び１０８は、データベース、ＣＤ－ＲＯＭ、テープ、デジタル記憶ライブラリのようなクエリーの１つ又はそれよりも多くのセットの検索及び記憶のために供給することができるデータベース又は記憶構造のあらゆる他のタイプとして実施することができる。データ記憶装置１０２及び１０８に保持されるクエリーは、複数の書記体系を有する所定の言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーを含むことができる。例えば、データ記憶装置１０２及び１０８に保持されるクエリーは、日本語の漢字、ひらがな、カタカナ、ＪＡＳＣＩＩ、及びローマ字書記体系に従って書かれたクエリーを含むことができる。

【００２６】

本発明の別の実施形態によると、検索エンジン１０７によって識別された関連するクエリーの候補セットは、１つ又はそれよりも多くのクエリーログにおいて統計的有意性で共起するクエリーの１つ又はそれよりも多くの連続する対を含む。検索エンジン１０７は、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃから受け取られたクエリーに関連する１つ又はそれよりも多くのクエリーを含む候補セットを識別するためにクエリーログを利用することができる。ユーザによって検索プロバイダ１００に提出された複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれる複数のクエリーは、クエリーログ構成要素１０６に保持することができる。クエリーログ構成要素１０６は、１つ又はそれよりも多くの書記体系に従って書かれた１つ又はそれよりも多くのク

エリーの記憶のために供給することができるデータベース又は類似の記憶構造として実施することができる。

【００２７】

クエリーログ構成要素１０６は、クエリーが検索プロバイダ１００に提出された頻度を識別する情報を保持することができる。同様に、クエリーログ構成要素１０６は、所定のクエリーが関連するクエリーに続く頻度を識別する情報を保持することができる。例えば、所定のセッション中、検索を行うユーザは、複数の書記体系を有する言語、例えば、日本語の１つ又はそれよりも多くの書記体系に従って書かれた「知的財産」という用語を含むクエリーを提出することができる。同じセッション中、ユーザは、１つ又はそれよりも多くの日本語書記体系に従って書かれた「特許弁理士」という用語を含むクエリーを提出することができる。クエリーログ構成要素１０６は、所定のユーザのセッション中にクエリー「特許弁理士」がクエリー「知的財産」に続く頻度を識別する情報を保持することができる。

【００２８】

検索エンジン１０７は、所定のクライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃから受け取られたクエリーに統計的に深く関連する１つ又はそれよりも多くのクエリーを含む候補セットを識別するためにクエリーログ構成要素１０６によって保持されたクエリーログを利用することができる。所定のクエリーに関連すると識別された１つ又はそれよりも多くのクエリーは、クエリーログ構成要素１０６に保持されたクエリーログで示すように、関連するクエリーの候補セットを補足するか又は生成するために使用することができる。関連するクエリーの候補セットは、日本語のような複数の書記体系を有する所定の言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーを含むことができる。クエリーログを使用して所定のクエリーに関連する１つ又はそれよりも多くのクエリーを識別するための例示的な方法は、「代替検索クエリーを判断するためのシステム及び方法」という名称の共同所有の米国特許出願出願番号第１１／２００、８５１号、及び「モジュラー最適化動的セット」という名称の米国特許仮出願第６０／７３６、１３３号に説明されており、これらの開示は、本明細書においてその全内容が引用により組み込まれている。

【００２９】

類似性構成要素１０４は、検索エンジン１０７によって識別された候補セットを使用して、関連するクエリーの候補セット内の１つ又はそれよりも多くのクエリーに対する類似性スコアを計算する。類似性構成要素１０４は、関連するクエリーの候補セットから所定のクエリーＱ’を選択し、所定のクライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃから受け取られた所定のクエリーＱに対するＱ’の意味における類似性の強さを示すＱ’の類似性スコアを計算するように作動する。類似性構成要素１０４は、本明細書に説明される方法に従って検索エンジン１０７によって識別された関連するクエリーの候補セット内の１つ又はそれよりも多くのクエリーの各々に対して類似性スコアを計算するように作動する。

【００３０】

類似性構成要素１０４は、検索エンジン１０７によって識別された関連するクエリーの候補セットにおける各クエリーＱ’に対する類似性スコアを計算するために変換構成要素１１０を利用することができる。本発明の一実施形態によると、変換構成要素１１０は、所定のクエリーを１つ又はそれよりも多くの文書フォーマットに変換する。変換構成要素１１０によって生成された所定のクエリーＱ’の１つ又はそれよりも多くの文書フォーマットは、類似性スコアの計算を容易にするために類似性構成要素１０４に分配することができる。例えば、類似性構成要素１０４は、正確な類似性スコアを計算するために、ユーザから受け取られた所定のクエリーＱと、関連するクエリーの候補セットから選択された関連するクエリーＱ’との多数の比較を行うことができる。しかし、上述のように、関連するクエリーの候補セット内の１つ又はそれよりも多くのクエリーは、複数の書記体系を有する所定の言語の１つ又はそれよりも多くの書記体系に従って書くことができる。同様

に、所定のクライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃから受け取られたクエリーは、複数の書記体系を有する所定の言語の１つ又はそれよりも多くの書記体系に従って書くことができる。類似性構成要素１０４によって行われる１つ又はそれよりも多くの比較は、ユーザから受け取られたクエリーＱと、関連するクエリーの候補セットから選択された所定のクエリーＱ’とを特定の書記体系に従って表現することができるように要求することができる。例えば、類似性構成要素１０４は、２つのクエリーを比較するために、所定のクエリーＱ及び関連するクエリーＱ’の１つ又はそれよりも多くのＪＡＳＣＩＩ文字をＡＳＣＩＩ文字に変換するように要求することができる。

【００３１】

様々な書記体系に従って書かれる可能なクエリーＱとクエリーＱ’を比較するために、類似性構成要素１０４は、所定のクエリーを変換構成要素１１０に分配することができる。本発明の一実施形態によると、変換構成要素１１０は、所定のクエリーに関連付けられた言語と書記体系を識別し、クエリーを１つ又はそれよりも多くの代替文書フォーマットに変換するように作動する。検索エンジン１０７によって識別された候補セットは、日本語の漢字、かな、ＪＡＳＣＩＩ、及びローマ字書記体系のような複数の書記体系を有する所定の言語の広範な書記体系に従って書かれたクエリーを含むことができる。変換構成要素１１０は、１つ又はそれよりも多くの日本語書記体系に従って書かれたクエリーを識別し、クエリーを１つ又はそれよりも多くの代替書記体系に変換するように作動する。例えば、変換構成要素１１０は、日本語のカタカナ書記体系に従って書かれたクエリーを識別し、日本語ローマ字書記体系に従ってクエリーを変換するように作動する。同様に、変換構成要素１１０は、１つ又はそれよりも多くのＪＡＳＣＩＩ文字を含むクエリーを識別し、類似性構成要素１０４による類似性スコアの計算を容易にするために１つ又はそれよりも多くのＪＡＳＣＩＩ文字をＡＳＣＩＩ文字に変換するように作動する。

【００３２】

本発明の一実施形態によると、関連するクエリーの候補セット内の１つ又はそれよりも多くのクエリーに対する類似性構成要素１０４によって計算された類似性スコアは、分配のために候補セットから１つ又はそれよりも多くのクエリーを選択するために分配構成要素１１６によって使用される。類似性スコアに基づくクエリーの選択は、所定のクエリーＱに対して意味において最も類似のクエリーの選択を可能にする。例えば、分配構成要素１１６は、所定の閾値を超える類似性スコアを有する１つ又はそれよりも多くのクエリーを関連するクエリーの候補セットから選択することができる。同様に、分配構成要素は、最も高い類似性スコアを有するＮ個のクエリーを候補セットから選択することができる。当業者は、類似性スコアを使用して候補セットから１つ又はそれよりも多くのクエリーを選択するための他の技術を認識する。

【００３３】

分配構成要素１１６は、候補セットから選択された１つ又はそれよりも多くのクエリーを分配することができる。本発明の一実施形態によると、分配構成要素１１６は、「示唆される代替クエリー」又は「意味において類似のクエリー」として候補セットから選択されたクエリーをユーザにネットワーク１２２を通じて表示する。代替的に又は上述のものと共に、分配構成要素１１６は、ネットワーク１２２に通信することができるように連結されたクライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃの所定のユーザによってビューされる検索結果ウェブページに選択されたクエリーを組み込むことができる検索エンジン１０７に選択された１つ又はそれよりも多くのクエリーを分配するように作動する。

【００３４】

候補セット内の１つ又はそれよりも多くのクエリーに対して類似性構成要素１０４によって計算された類似性スコアは、所定の要求に応じた分配のための広告を含むコンテンツの１つ又はそれよりも多くの項目を選択するために更に使用することができる。本発明の一実施形態によると、広告は、上述のデータ記憶装置１０２及び１０８、又は１つ又はそれよりも多くの異なるデータ記憶装置（示されない）に保持することができる。１つ又は

それよりも多くのローカル１０２、リモート１０８、又は異なるデータ記憶装置は、１つ又はそれよりも多くの広告及び広告に対応する語に対する関連の指し値を保持するように作動する。例えば、所定の広告主が、ノートブックコンピュータに対する所定の広告の表示を望むとする。従って、広告主は、「ノートブックコンピュータ」という用語に対して入札し、「ノートブックコンピュータ」という用語を含むクエリーに応答して表示される広告を識別することができる。検索プロバイダ１００がクエリーを受け取った場合、検索エンジン１０７は、ローカル及びリモートデータ記憶装置１０２及び１０８、又は１つ又はそれよりも多くの異なるデータ記憶装置を検索し、１つ又はそれよりも多くの広告主が受け取られたクエリーを含む１つ又はそれよりも多くの用語に対して入札したか否かを判断することができる。クエリーを含む用語に対する１つ又はそれよりも多くの指し値が識別された場合、１つ又はそれよりも多くの用語に対する指し値に関連付けられた広告が検索され、分配構成要素１１６を使用してユーザのクライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃ上でユーザに表示される。ユーザが表示された所定の広告を選択した場合、選択された広告に関連付けられた広告主に、広告主の指し値に従って合計額が請求される。

【００３５】

しかし、広告主は、複数の書記体系を有する言語の単に１つの書記体系に従って書かれた用語への入札を選択することができる。例えば、広告主は、日本語のひらがな書記体系だけに従って書かれた用語への入札を選択することができる。しかし、上述のように、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃのユーザによって提出された１つ又はそれよりも多くの検索クエリーは、１つ又はそれよりも多くの書記体系に従って書かれた用語及び語句を含むことができる。従って、検索エンジン１０７は、所定のクエリーに応答して検索される広告の幅を拡大するために所定の閾値を超える類似性スコアを有するクエリーを利用することができる。本発明の一実施形態によると、検索エンジン１０７は、所定の閾値を超える類似性スコアを有する１つ又はそれよりも多くのクエリーを含む用語に応答して１つ又はそれよりも多くの広告を識別する。所定の閾値を超える類似性スコアを有するクエリーを含む用語に応じるとして識別された１つ又はそれよりも多くの広告を、１つ又はそれよりも多くのクライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃへの分配のために選択することができる。

【００３６】

例えば、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃのユーザは、日本語の漢字及びローマ字書記体系の両方に従って書かれた日本語の語から成る検索クエリーＱを作成することができる。ユーザは、ネットワーク１２２を通じて検索プロバイダ１００にクエリーを提出することができる。検索エンジン１０７は、ユーザによって用いられた漢字及びローマ字語に対して入札した広告主がないと判断することができる。代替的に又は上述の事柄と共に、検索エンジン１０７は、ユーザによって用いられた漢字及びローマ字語に関連付けられた指し値に対応する広告を表示することは殆ど収益をもたらさないと判断することができる。しかし、検索エンジン１０７は、関連した指し値を有する１つ又はそれよりも多くの用語を識別するために、所定の閾値を超える類似性スコアを有する候補セットから選択された１つ又はそれよりも多くのクエリーを含む用語を利用することができる。同様に、検索エンジン１０７は、所定の閾値を超える指し値を有する１つ又はそれよりも多くの用語を識別するために、所定の閾値を超える類似性スコアを有する候補セットから選択された１つ又はそれよりも多くのクエリーを含む用語を利用することができる。検索エンジン１０７は、その後、ユーザによって作成された検索クエリーＱに応答して１つ又はそれよりも多くの広告を選択するために、関連の指し値を有する１つ又はそれよりも多くの用語、又は所定の閾値を超える関連の指し値を有する１つ又はそれよりも多くの用語を利用することができる。

【００３７】

別の実施例によると、所定の閾値を超える類似性スコアを有する候補セットから選択された所定のクエリーＱ’がひらがな用語を含むとすると、ユーザによって作成された上述

のクエリーＱは、漢字とローマ字語を含む。検索エンジンは、１つ又はそれよりも多くの広告主がクエリーＱ’を含むひらがな用語に入札したか否かを判断するために、クエリーＱ’を含む１つ又はそれ上のひらがな用語を利用することができる。同様に、検索エンジンは、１つ又はそれよりも多くの広告主が、所定の閾値を超えるクエリーＱ’を含む１つ又はそれよりも多くのひらがな用語に入札したか否かを判断することができる。検索エンジン１０７は、クエリーＱ’を含む用語に対して関連の指し値を有する１つ又はそれよりも多くの広告を検索し、１つ又はそれよりも多くの広告を分配構成要素に分配することができる。本発明の一実施形態によると、検索エンジン１０７は、クエリーＱ’を含む１つ又はそれよりも多くの用語に対して最も関連のある指し値を有する１つ又はそれよりも多くの広告を検索する。分配構成要素１１６は、その後、クエリーＱを提出したユーザに１つ又はそれよりも多くの広告を分配することができる。

【００３８】

上述の実施形態は、クエリーの受け取り及び処理を示しているが、図１に示されている検索プロバイダ１００システムは、クエリーに対する類似性スコアの受け取り及び計算に制限されず、テキストの１つ又はそれよりも多くのストリングを含む１つ又はそれよりも多くの用語に対する類似性スコアを計算するために更に使用することができる。クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃのユーザは、検索プロバイダ１００に、限定ではないが、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた語句、文、段落、及び文書を含む１つ又はそれよりも多くの用語を含むテキストの１つ又はそれよりも多くのストリングを分配することができる。従って、検索プロバイダ１００は、テキストのこれらの１つ又はそれよりも多くのストリングのログを１つ又はそれよりも多くのログファイルに記録する。検索プロバイダ１００は、このログファイルから１つ又はそれよりも多くの項目を含む候補セットを識別するように作動可能であり、ここで、所定の項目は、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃの所定のユーザによって分配された１つ又はそれよりも多くの用語に関連する用語の１つ又はそれよりも多くのセットを含む。例えば、候補セットの所定の項目は、語句又は文を含むことができる。同様に、候補セットの所定の項目は、段落又は全文書を含むことができる。検索プロバイダは、クライアントデバイス１２４ａ、１２４ｂ、及び１２４ｃから受け取られた１つ又はそれよりも多くの用語に対して項目の意味における類似性の強さを示す候補セットの１つ又はそれよりも多くの項目に対する類似性スコアを計算することができる。

【００３９】

図２は、所定のクエリーＱに意味において関連する１つ又はそれよりも多くのクエリーＱ’を候補セットから選択する方法の一実施形態を示し、ここで、クエリーＱ及びクエリーＱ’は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれる。図２に示すように、検索クエリーが所定のユーザから受け取られる（段階２０５）。クエリーは、「インターネット」のようなネットワークに通信することができるように連結されたクライアントデバイスから受け取られ、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系の組合せに従って書かれた１つ又はそれよりも多くの用語又は語句を含むことができる。例えば、ユーザから受け取られたクエリーは、漢字、カタカナ、及びひらがな書記体系に従って書かれた日本語の語を含むことができる。

【００４０】

ユーザによって作成された所定のクエリーＱに関連するクエリーから成る候補セットが識別される（段階２１０）。候補セットは、ユーザのクエリーに関連付けられた言語の１つ又はそれよりも多くの書記体系に従って書かれたクエリーから構成することができる。例えば、所定のクエリーＱは、クエリー「ラクテン」のような日本語のカタカナ書記体系に従って書かれた用語を含むことができる。従って、関連するクエリーの候補セットは、１つ又はそれよりも多くの日本語書記体系の１つ又はそれよりも多くの組合せに従って書かれた１つ又はそれよりも多くのクエリーを含むことができる。例えば、上述のひらがなクエリー「ラクテン」に関連するクエリーの候補セットは、ローマ字クエリー「ｒａｋｕ

ｔｅｎ」、漢字クエリー「楽天」、ひらがなクエリー「らくてん」などを含むことができる。

【００４１】

所定のクエリーＱに関連するクエリーの候補セットは、１つ又はそれよりも多くのクエリーログを使用して生成することができる。本発明の一実施形態によると、クエリーログは、所定のクエリーセッション中にユーザによって作成された１つ又はそれよりも多くのクエリーを識別することができる。例えば、所定のクエリーセッション中、ユーザは、日本語のひらがな及び漢字書記体系に従って書かれた用語を含むクエリーを作成することができる。同じクエリーセッション中、ユーザは、日本語のカタカナ及びローマ字書記体系に従って書かれた用語を含むクエリーを作成することができる。２つのクエリーが統計的有意性で１つ又はそれよりも多くのクエリーログに共起するか否かを判断するための分析を行うことができる。本発明の一実施形態によると、統計的有意性閾値は、１つ又はそれよりも多くのクエリーログで示すような所定のクエリーＱに最も関連する１つ又はそれよりも多くのクエリーを選択するために使用することができる。

【００４２】

候補セットは、統計的有意性、又は１つ又はそれよりも多くのクエリーログで示すような所定の閾値を超える統計的有意性で所定のクエリーに関連するとして識別された１つ又はそれよりも多くのクエリーで生成することができる。関連するクエリーの候補セットを含む１つ又はそれよりも多くのクエリーは、全体が引用により組み込まれている上述の出願に説明されるクエリーログを使用して統計的有意性で関連するクエリーを判断する方法に従って選択される。

【００４３】

所定のクエリーＱ’は、関連するクエリーの候補セットから選択される（段階２１５）。図２に示す実施形態によると、類似性スコアは、選択されたクエリーＱ’に対して計算される（段階２２０）。所定のクエリーＱ’に対して計算された類似性スコアは、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた所定のクエリーＱの意味に対するクエリーＱ’の意味の類似性の強さを示す数値を提供する。表Ａは、所定のクエリーＱ’に対する類似性スコアを計算するために使用することができる方程式の一実施形態を示している。

【００４４】

表Ａに示す式は、限定ではないが、漢字、かな、ＪＡＳＣＩＩ、かな、カタカナ、ローマ字、及びひらがなを含む１つ又はそれよりも多くの日本語書記体系に従って書かれる所定のクエリーＱに対する所定のクエリーＱ’の意味における類似性の強さを示すスコアを計算するために使用することができる。当業者は、複数の書記体系を有する他の言語に対する類似性スコアの計算を提供することができるように表Ａに示されている式を修正することができることを認識する。

【００４５】

（表Ａ）

| |
|---|
| 類似性スコア（Ｑ’）＝１．４７５５１＋ｌｅｖｒ（Ｑ、Ｑ’）ｘ?１．６８８２１＋ｌｅｖｒｓ（Ｑ、Ｑ’）ｘ２．４８７００＋ｗｏｒｄｒ（Ｑ、Ｑ’）ｘ０．４４３６６＋数字（Ｑ、Ｑ’）ｘ０．７５３８８＋ｋａｎｊｉｄ（Ｑ、Ｑ’）ｘ０．２２４９６＋ｏｐｒ（Ｑ、Ｑ’）ｘ?０．４００８３＋日本語（Ｑ、Ｑ’）ｘ０．０９３６８＋ｌｅｖｋ（Ｑ、Ｑ’）ｘ?０．３２５７４＋ｐ１２ｍｉｎ（Ｑ、Ｑ’）ｘ?０．３３２５８ |

【００４６】

表Ａに示す式によると、Ｑは、１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーを表している。Ｑ’は、クエリーＱに関連するクエリーの候補セットから選択されたクエリーを表している。Ｌｅｖｒは、全ての日本語文字をローマ字に変換

した後にＱとＱ’間の文字編集距離を計算するための関数である。ｌｅｖｒｓは、全ての日本語文字をローマ字に変換しスペースを取り除いた後にＱとＱ’間の文字編集距離を計算するための関数である。Ｗｏｒｄｒは、全ての日本語文字をローマ字に変換した後のＱとＱ’間の語編集距離である。数字は、ＱがＱ’に現れないどの数字も包含するか否かを及び逆も同様か否かを識別するための関数である。Ｋａｎｊｉｉｄは、Ｑ又はＱ’のいずれかが漢字文字を包含するか否かを判断するための関数であり、包含する場合、ＱとＱ’間の漢字相違を識別するための関数である。Ｏｐｒは、各クエリーの全ての日本語文字がローマ字に変換された後、第１文字の不一致まで各クエリーの最左端文字から始まるＱ及びＱ’が共通して有する文字の数を計算し、計算を継続させるための関数である。ｌｅｖｋは、全ての漢字文字がかな文字に変換され全ての非日本語文字が取り除かれた後に、ＱとＱ’間の文字編集距離を計算するための関数である。Ｐ１２ｍｉｎは、ユーザクエリーセッションのログにおいてクエリーＱに続くクエリーＱ’のクエリー置換確率を計算するための関数である。表Ａに示す類似性スコア関数によって利用される関数の実施形態を図３から図１１に示している。

【００４７】

類似性スコアが候補セット内の１つ又はそれよりも多くのクエリーに対して計算されたか否かを判断するための検査が行われる（段階２２５）。候補セット内の１つ又はそれよりも多くのクエリーが関連する類似性スコアを持たない場合、付加的なクエリーＱ’が、候補セットから選択される（段階２１５）。代替的に、類似性スコアが候補セット内の１つ又はそれよりも多くのクエリーに対して計算されている場合、所定のクエリーＱ’が、候補セットから選択される（段階２３０）。候補セットから選択されたクエリーＱ’に関連した類似性スコアが、所定の類似性スコア閾値を超えるか否かを判断するための検査が行われる（段階２３５）。本発明の一実施形態によると、類似性スコア閾値は、所定のクエリーＱ’に関連付けられた類似性スコアとの比較を行うために使用することができる数値を含む。類似性スコアは、クエリーＱに対する所定のクエリーＱ’の意味における類似性の強さを示すので、類似性スコア閾値の使用は、クエリーＱに対して意味において最も類似する１つ又はそれよりも多くのクエリーの候補セットからの選択を容易にする。

【００４８】

所定のクエリーＱ’に関連付けられた類似性スコアが、類似性スコア閾値を超える場合、クエリーＱ’が分配セットに加えられる（段階２４５）。本発明の一実施形態によると、分配セットは、類似性スコア閾値を超える類似性スコアを有する候補セットから選択された１つ又はそれよりも多くのクエリーを含む。所定のクエリーＱ’に関連付けられた類似性スコアが、類似性スコア閾値を超えない場合、クエリーＱ’は、分配セットに加えられない（段階２４０）。

【００４９】

分析を必要とする候補セットの付加的なクエリーがあるか否かを判断するための検査が行われる（段階２５０）。候補の１つ又はそれよりも多くのクエリーが分析を必要とする場合、付加的なクエリーＱ’が、候補セットから選択される（段階２３０）。代替的に、候補セットの全てのクエリーが分析され、分配セットに類似性スコア閾値を超える１つ又はそれよりも多くのクエリーがポピュレートされた後、分配セットの１つ又はそれよりも多くのクエリーが分配される（段階２５５）。

【００５０】

類似性スコア閾値を超えるクエリーの分配セットの１つ又はそれよりも多くのクエリーが、クエリーＱを提出したユーザに分配される。本発明の一実施形態によると、分配セットの１つ又はそれよりも多くのクエリーが、結果ウェブページでユーザに表示される。例えば、ユーザには、クエリーＱに応じたコンテンツ項目へのリンク、並びにクエリーＱに対する意味において最も類似の分配セットを含む１つ又はそれよりも多くのＱ’クエリーのような結果を含むウェブページが表示される。所定のユーザに分配された分配セットの１つ又はそれよりも多くのクエリーは、クエリーＱに対する所定のクエリーＱ’の意味における類似性の相対的な強さをユーザに示すために、類似性スコアに従ってランク付けさ

れたリストで表示される。

【００５１】

図３から１１は、クエリーの候補セットから選択された所定のクエリーＱ’に対する類似性スコアを計算するために使用することができる表Ａに示す関数の実施形態を示している。上述のように、表Ａ、更に図３から１１に示す複数の関数は、１つ又はそれよりも多くの日本語書記体系に従って書かれたクエリーＱに対する所定のクエリーＱ’の意味における類似性の強さを示す類似性スコアを計算するために使用することができる。しかし、当業者は、図３から１１に示す関数の実施形態が例示的なものであり、日本語言語及び書記体系に制限されないものであること、及び複数の書記体系を有する他の言語に対する類似性スコアの計算を提供するように修正することができることを認識する。当業者は、更に、図３から１１に示されている関数が、所定のクエリーに関連する１つ又はそれよりも多くのクエリーを含む候補セットに対する類似性スコアを計算することに制限されないこと、及び複数の技術に従って選択された１つ又はそれよりも多くのクエリーを含むクエリーの候補セットに対する類似性スコアを計算するために使用することができることを認識する。更に、当業者は、図３から１１に示す関数が、１つ又はそれよりも多くのクエリーを含む候補セットに対する類似性スコアを計算することに制限されないこと、更に、限定ではないが、語句、文、段落、及び文書を含む用語の１つ又はそれよりも多くのセットに対する類似性スコアを計算するために修正することができることを認識する。

【００５２】

図３は、１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱと、クエリーの候補セットから選択されたクエリーＱ’との間の文字編集距離を計算する方法の一実施形態を示している。図３に示す方法は、表Ａに示す類似性スコア関数によって利用されるｌｅｖｋ関数の一実施形態を示している。

漢字、カタカナ、ひらがなのような１つ又はそれよりも多くの日本語書記体系に従って書かれるクエリーＱを含む１つ又はそれよりも多くの文字がローマ字に変換される（段階３０５）。所定のクエリーＱ’は、１つ又はそれよりも多くのクエリーから構成される候補セットから選択される（段階３１０）。候補セットから選択されたクエリーＱ’は、クエリーＱに関連付けられた言語の１つ又はそれよりも多くの書記体系に従って書くことができる。例えば、Ｑ’は、クエリーＱと同じ書記体系、又は日本語ローマ字書記体系、日本語かな書記体系のような１つ又はそれよりも多くの代替日本語書記体系に従って書くことができる。Ｑ’を含む文字がローマ字形式であるか否かを判断するための検査が行われる（段階３１５）。クエリーＱ’がローマ字形式でない場合、Ｑ’を含む１つ又はそれよりも多くの文字がローマ字に変換される（段階３２０）。Ｑ’を含む１つ又はそれよりも多くの用語が既にローマ字形式である場合、又はＱ’の文字全てがローマ字形式に変換された後に、クエリーＱとクエリーＱ’間の文字編集距離を識別するための計算が行われる（段階３２５）。文字編集距離値は、Ｑ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数に供給される。

【００５３】

図４は、１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱと、クエリーの候補セットから選択されたクエリーＱ’との間の文字編集距離を計算する方法の一実施形態を示している。図４に示す実施形態は、表Ａに示す類似性スコア関数によって使用されるｌｅｖｒｓ関数の一実施形態を提供する。

図４に示す実施形態によると、漢字、カタカナ、又はひらがなのような１つ又はそれよりも多くの日本語書記体系に従って書かれたクエリーＱがローマ字形式に変換される（段階４０５）。その後、クエリーＱからローマ字で現れる全てのスペース文字が取り除かれる（段階４０８）。例えば、所定のクエリーＱは、漢字の用語「電車男」を含むことができる。ローマ字形式に変換後、クエリーＱは、用語「ｄｅｎｓｈａ　ｏｔｏｋｏ」を含むことができ、スペースを取り除いた後、クエリーＱは、文字「ｄｅｎｓｈａｏｔｏｋｏ」を含むことができる。

【００５４】

所定のクエリーＱ’が１つ又はそれよりも多くのクエリーを含む候補セットから選択される（段階４１０）。Ｑ’がローマ字形式であるか否かを判断するための検査が行われる（段階４１５）。クエリーＱ’がローマ字形式でない場合、クエリーＱ’を含む１つ又はそれよりも多くの文字がローマ字に変換される（段階４２０）。クエリーＱ’を含む文字が既にローマ字形式である場合、又はクエリーＱ’を含む文字がローマ字形式に変換された後に、クエリーＱ’内の全てのスペースが取り除かれる（段階４２５）。その後、クエリーＱとＱ’のローマ字形式間の文字編集距離が計算される（段階４３０）。クエリーＱとＱ’間の計算された文字編集距離は、Ｑ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数によって使用される。

【００５５】

図５は、表Ａに示すｗｏｒｄｒ関数の一実施形態を示している。図５に示すｗｏｒｄｒ関数の実施形態は、１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱとクエリーの候補セットから選択されたクエリーＱ’との間の語編集距離の計算を提供する。本発明の一実施形態によると、所定のクエリーＱとクエリーＱ’間の語編集距離は、値１と、ＱとＱ’におけるスペースで区切られた固有の共起語の数とＱとＱ’の両方におけるスペースで区切られた固有の語の総数との商との間の差である。

【００５６】

１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱを含む文字がローマ字形式に変換される（段階５０５）。その後、所定のクエリーＱ’がクエリーの候補セットから選択される（段階５０６）。クエリーＱ’がローマ字形式であるか否かを判断するための検査が行われる（段階５０８）。クエリーＱ’がローマ字形式でない場合、クエリーＱ’を含む文字がローマ字に変換される（段階５１０）。クエリーＱ’を含む文字が既にローマ字形式である場合、又はＱ’を含む文字がローマ字形式に変換された後に、Ｑ及びＱ’におけるスペースで区切られた固有の共起語の数が識別される（段階５１５）。Ｑ及びＱ’におけるスペースで区切られた固有の共起語の数とＱ及びＱ’両方におけるスペースで区切られた固有の語の総数との商が計算される（段階５２０）。本発明の一実施形態によると、スペースで区切られた固有の共起語の数は、所定のクエリーＱ及びクエリーＱ’の両方に現れる固有の語の数を含む。更に、Ｑ及びＱ’の両方におけるスペースで区切られた固有の語の総数は、所定のクエリーＱ及びクエリーＱ’におけるスペースで区切られた固有の語の和を含む。

【００５７】

値１と、計算された商との間の差が計算され（段階５２５）、「ｗｏｒｄｒ」レジスタに割り当てられる（段階５３０）。本発明の一実施形態によると、「ｗｏｒｄｒ」レジスタは、所定の数値を記憶するためのメモリデバイスを含む。「ｗｏｒｄｒ」レジスタに割り当てられた値は、クエリーＱ’に対する類似性スコアを計算するために、表Ａに示されている類似性スコア関数によって使用される。

例えば、ローマ字形式の所定のクエリーＱは、用語「ｋｕｒｕｍａ　ｋｅｍｕｒｉ」から構成される。同様に、ローマ字形式の所定のクエリーＱ’は、用語「ｓｏｒａ　ｋｅｍｕｒｉ」から構成される。Ｑ及びＱ’におけるスペースで区切られた固有の共起語の数は、１、すなわち、語「ｋｅｍｕｒｉ」であり、ここで、Ｑ及びＱ’両方におけるスペースで区切られた固有の語の総数は、３、すなわち、語「ｋｕｒｕｍａ」、「ｓｏｒａ」、及び「ｋｅｍｕｒｉ」である。従って、Ｑ及びＱ’におけるスペースで区切られた固有の共起語の数と、Ｑ及びＱ’両方におけるスペースで区切られた固有の共起語の総数との商は、１／３である。更に、１と計算された商との間の差は、２／３である。値２／３は、「ｗｏｒｄｒ」レジスタに割り当てられ、クエリーＱ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数によって使用される。

【００５８】

図６は、クエリーの候補セットから選択されたクエリーＱ’との比較において、数字が１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱに固有であるか否かを判断する方法の一実施形態を示している。図６に示す実施形態は、表Ａに示す

類似性スコア関数によって使用される「数字」関数の一実施形態を提供する。

所定のクエリーＱ’は、１つ又はそれよりも多くの書記体系に従って書かれたクエリーから構成される候補セットから選択される（段階６０５）。所定のクエリーＱにおける数字が、クエリーＱ’に現れないか否かを判断するための検査が行われる。例えば、所定のクエリーＱは、日本語漢数字「六十八」（アラビア数字「６８」によって表される値に対応する）を包含することができ、所定のクエリーＱ’は、日本語漢数字「九十八」（アラビア数字「９８」によって表される値に対応する）を包含することができる。従って、段階６１０で行われる検査は、日本語漢数字「六」がクエリーＱ’に現れない場合、日本語漢数字「六」がクエリーＱに固有であると判断する。同様に、所定のクエリーＱは、日本語漢字文字とアラビア数字の「楽天２００５」を含むことができ、所定のクエリーＱ’は、日本語漢字文字とアラビア数字の「楽天２００４」を含むことができる。段階６１０で行われる検査は、アラビア数字５がクエリーＱ’に現れない場合、アラビア数字５がクエリーＱに固有であると判断する。

【００５９】

数字がクエリーＱに現われて、クエリーＱ’に現われないと識別された場合、「数字」レジスタは、クエリーＱが、クエリーＱ’にない数字を包含することを示す値１に設定される（段階６２０）。本発明の一実施形態によると、「数字」レジスタは、所定の数値を記憶するためのメモリデバイスを含む。

代替的に、Ｑ’が、クエリーＱに現われる１つ又はそれよりも多くの数字の各々を包含する場合、クエリーＱ’の数字がクエリーＱに現われないか否かを判断するための付加的な検査が行われる（段階６１５）。クエリーＱ’が、クエリーＱに現われない数字を包含する場合、上述の「数字」レジスタは、クエリーＱ’が、Ｑ’に固有の数字を包含することを示す値１に設定される（段階６２０）。代替的に、クエリーＱがＱ’における１つ又はそれよりも多くの数字の各々を包含する場合、「数字」レジスタは、クエリーＱ’における１つ又はそれよりも多くの数字がクエリーＱに現われること及び逆も同様に示す０に設定される（段階６２５）。「数字」レジスタに割り当てられる値、０又は１のいずれかは、クエリーＱ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数によって使用される。

【００６０】

図７は、表Ａに示す類似性スコア関数によって使用される「ｋａｎｊｉｄ」関数の一実施形態を示している。１つ又はそれよりも多くの日本語書記体系に従って書かれる所定のクエリーＱが受け取られる（段階７０５）。クエリーＱが１つ又はそれよりも多くの日本語漢字文字を包含するか否かを判断するための検査が行われる（段階７１０）。クエリーＱがいずれの漢字文字も包含しない場合、「ｋａｎｊｉｄ」レジスタは、０に設定され（段階７０８）、ここで、「ｋａｎｊｉｄ」レジスタは、所定の数値を記憶するためのメモリデバイスを含むことができる。代替的に、クエリーＱが１つ又はそれよりも多くの漢字文字を包含する場合、クエリーＱ’が、クエリーの候補セットから選択される（段階７１５）。

【００６１】

候補セットから選択されたクエリーＱ’が、１つ又はそれよりも多くの漢字文字を包含するか否かを判断するための検査が行われる（段階７２０）。クエリーＱ’がいずれの漢字文字も包含しない場合、上述の「ｋａｎｊｉｄ」レジスタは、０に設定される（段階７０８）。対照的に、Ｑ’が１つ又はそれよりも多くの漢字文字を包含する場合、Ｑ及びＱ’における１つ又はそれよりも多くの漢字でない文字が取り除かれる（段階７２２）。その後、クエリーＱ及びクエリーＱ’に共起する固有の漢字文字の数が識別される（段階７２５）。例えば、漢字でない文字を取り除いた後、クエリーＱが、漢字文字「楽天市場」から構成され、漢字でない文字を取り除いた後、クエリーＱ’が、漢字文字「楽天」から構成される場合、Ｑ及びＱ’における固有の共起漢字文字の数は、２、すなわち、「楽天」である。

【００６２】

その後、Ｑ及びＱ’両方における固有の漢字文字の総数が識別される（段階７２７）。例えば、漢字文字「楽天市場」から構成されるＱ及び漢字文字「楽天」から構成されるＱ’両方における固有の漢字文字の総数は、６、すなわち、クエリーＱからの固有の漢字文字「楽天市場」とクエリーＱ’からの固有の漢字文字「楽天」である。共起する漢字文字の数と総固有漢字文字との商が計算される（段階７３０）。「ｋａｎｊｉｄ」レジスタは、１と計算された商との間の差の値に設定される（段階７３５）。「ｋａｎｊｉｄ」レジスタ値は、Ｑ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数によって使用される。

【００６３】

図８は、１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱと、クエリーの候補セットから選択されたクエリーＱ’との接頭辞において重なる文字の数を識別し、更に、第１文字の不一致まで各クエリーの最左端の文字の比較から始めて比較を継続させる方法の一実施形態を示している。図８に示す方法は、表Ａに示す類似性スコア関数によって利用されるｏｐｒ関数の一実施形態を示している。

【００６４】

１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱが、ローマ字形式に変換される（段階８０５）。クエリーＱ’が、クエリーの候補セットから選択される（段階８１０）。クエリーＱ’を含む１つ又はそれよりも多くの文字が、ローマ字形式であるか否かを判断するための検査が行われる（段階８１５）。クエリーＱ’を含む１つ又はそれよりも多くの文字がローマ字形式でない場合、文字が、ローマ字に変換される（段階８２０）。Ｑ’を含む文字が既にローマ字形式である場合、又はＱ’を含む１つ又はそれよりも多くの文字がローマ字形式に変換された後に、クエリーＱ及びクエリーＱ’の第１ローマ字文字が選択される（段階８２５）。

【００６５】

クエリーＱから選択された第１文字とクエリーＱ’から選択された第１文字とが適合するか否かを判断するための検査が行われる（段階８３５）。Ｑ及びＱ’から選択された第１文字が適合しない場合、処理は終了する（段階８３０）。代替的に、選択された文字が適合した場合、クエリーＱ及びクエリーＱ’に対する文字適合が識別されたことを示す文字適合計数レジスタが増分される（段階８５０）。本発明の一実施形態によると、文字適合計数レジスタは、値０で初期化され、クエリーＱ及びクエリーＱ’からの文字が適合として識別された場合に増分される。

【００６６】

Ｑ及びＱ’からの次の文字が選択され（段階８４０）、次の文字が適合するか否かを判断するための検査が行われる（段階８３５）。Ｑ及びＱ’から選択された文字が適合しない場合、文字適合計数レジスタは増分されず、処理は終了する（段階８３０）。処理が終了した場合（段階８３０）、文字適合計数レジスタの値は、Ｑ及びＱ’において適合する文字の数を示すことになる。文字適合計数レジスタの値は、クエリーＱ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数によって利用される。

【００６７】

図９は、１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱ又はクエリーの候補セットから選択されたクエリーＱ’が非ローマ字文字を包含するか否かを識別する方法の一実施形態を示している。図９に示す実施形態は、表Ａに示されている類似性スコア関数によって使用される「日本語」関数を示している。

１つ又はそれよりも多くの日本語書記体系に従って書かれた所定のクエリーＱが受け取られる（段階９０５）。クエリーＱが１つ又はそれよりも多くの非ローマ字文字を包含するか否かを判断するための検査が行われる（段階９１０）。クエリーＱが１つ又はそれよりも多くの非ローマ字文字を包含する場合、「日本語」レジスタは、値１に設定される（段階９０８）。本発明の一実施形態によると、「日本語」レジスタは、所定の数値を記憶するためのメモリデバイスを含む。

【００６８】

クエリーＱが１つ又はそれよりも多くの非ローマ字文字を包含しない場合、クエリーＱ’が、１つ又はそれよりも多くのクエリーを含む候補セットから選択される（段階９１５）。クエリーＱ’が、１つ又はそれよりも多くの非ローマ字文字を包含するか否かを判断するための検査が行われる（段階９２０）。クエリーＱ’が、１つ又はそれよりも多くの非ローマ字文字を包含する場合、「日本語」レジスタは、値（「１」）に設定される（段階９０８）。代替的に、クエリーＱ’が非ローマ字文字だけを包含する場合、「日本語」レジスタは、値０に設定され（段階９２２）、その後、処理が終了する（段階９２５）。「日本語」レジスタに保持される値は、クエリーＱ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数によって利用される。

【００６９】

図１０は、全ての漢字及び非日本語文字が各それぞれのクエリーから取り除かれた後、所定のクエリーＱ及びクエリーＱ’の間の文字編集距離を判断する方法の一実施形態を示している。図１０に示されている方法は、表Ａに示す類似性スコア関数によって利用されるｌｅｖｋ関数の一実施形態を示している。

図１０に示すように、所定のクエリーＱ’が、クエリーの候補セットから選択される（段階１００５）。１つ又はそれよりも多くの日本語書記体系に従って書かれたクエリーＱ’又は所定のクエリーＱが、１つ又はそれよりも多くの漢字文字を包含するか否かを判断するための検査が行われる（段階１０１０）。クエリーＱ又はクエリーＱ’のいずれかが、１つ又はそれよりも多くの漢字文字を包含する場合、各それぞれのクエリーにおける漢字文字が、かな文字に変換される（段階１０１５）。例えば、クエリーＱは、「人２００」のような漢字文字とアラビア数字の両方から構成される。漢字文字をかな文字に変換した後、クエリーＱは、文字「ひと２００」を含むことができる。

【００７０】

クエリーＱ又はクエリーＱ’のいずれも漢字文字を包含しない場合、又は各それぞれのクエリーにおける全ての漢字文字がかな文字に変換された後に、いずれかのクエリーが非日本語文字を包含するか否かを判断するための検査が行われる（段階１０２０）。本発明の一実施形態によると、非日本語文字は、１つ又はそれよりも多くの日本語書記体系に従って書かれていない文字を含む。例えば、クエリーＱが、「ひと２００」のようなかな文字とアラビア数字を含む場合、アラビア数字「２００」は、非日本語文字を構成することができる。

【００７１】

クエリーＱ又はクエリーＱ’のいずれかが、非日本語文字を包含する場合、非日本語文字が取り除かれる（段階１０２５）。上述の実施例に関して、クエリーＱから非日本語文字、すなわち、アラビア数字「２００」を取り除いた後、クエリーＱは、かな文字「ひと」を含むことができる。クエリーＱ又はクエリーＱ’のいずれも非日本語文字を包含しない場合、又は全ての非日本語文字が取り除かれた後に、ＱとＱ’間の文字編集距離が計算される（段階１０３０）。クエリーＱとクエリーＱ’間の文字編集距離は、Ｑ’に対する類似性スコアを計算するために、表Ａに示す類似性スコア関数によって使用される。

【００７２】

図１１は、表Ａに示す類似性スコア関数によって利用される「ｐ１２ｍｉｎ」関数の一実施形態を示している。本発明の一実施形態によると、「ｐ１２ｍｉｎ」関数は、所定のクエリーＱに続く所定のクエリーＱ’のクエリー置換確率を計算し、所定の語句Ｐに続く語句Ｐ’の語句置換を計算するために使用される。例えば、１つ又はそれよりも多くのクエリーログは、クエリーセッション中に所定のユーザによって提出された１つ又はそれよりも多くのクエリー及び語句を識別する段階を保持することができる。クエリーログは、例えば、ユーザがクエリーＱをどのように精錬したか、ユーザがクエリーＱをどのように書き換えたか、クエリーＱを表すためにユーザが複数の書記体系を有する言語の１つ又はそれよりも多くの代替書記体系をどのように利用したかなどの指示を提供するために、ユーザによって提出された１つ又はそれよりも多くのクエリー及び語句の順序を識別することができる。クエリーログは、更に、１人又はそれよりも多くのユーザが、１つ又はそれ

よりも多くのクエリー又は語句を提出した頻度を指示することができる。

【００７３】

所定のクエリーＱが１つ又はそれよりも多くのクエリーログに現われる頻度が識別される（段階１１０５）。所定のクエリーＱ’が、クエリーの候補セットから選択される（段階１１１０）。１つ又はそれよりも多くのクエリーログのいずれかにおいてクエリーＱ’がクエリーＱに続くか否かを判断するための検査が行われる（段階１１１５）。本発明の一実施形態によると、所定のユーザのクエリーセッションに対してクエリーログにおいてクエリーＱ’がクエリーＱに続くか否かを判断するための検査が行われ、クエリーセッションは、所定の期間にユーザによって提出された１つ又はそれよりも多くのクエリーを含むことができる。

【００７４】

クエリーＱ’が、１つ又はそれよりも多くのクエリーログのいずれかにおいてクエリーＱに続かない場合、「ｐ１２ｍｉｎ」レジスタは、０に設定され（段階１１２５）、「ｐ１２ｍｉｎ」レジスタは、所定の数値を記憶するためのメモリデバイスを含むことができる。代替的に、クエリーＱ’が、クエリーログの１つ又はそれよりも多くにおいてＱに続くものとして識別された場合、クエリーＱ’がクエリーログにおいてクエリーＱに続く頻度が識別される（段階１１２０）。「ｐ１２ｍｉｎ」レジスタは、クエリーＱ’がクエリーログにおいてクエリーＱに続く頻度と、クエリーログにおけるクエリーＱの頻度との商の値に設定される（段階１１４０）。例えば、クエリーＱがクエリーログに１２回現われ、Ｑ’がクエリーログにおいてクエリーＱに７回続く場合、「ｐ１２ｍｉｎ」レジスタは、値「７／１２」に設定される。

【００７５】

当業者は、図３から１１に示し、かつ表Ａに示す類似性スコア関数によって利用される関数が日本語に制限されないこと、及び複数の書記体系を有する１つ又はそれよりも多くの言語に対して修正することができることを認識する。当業者は、更に、表Ａに示す類似性スコア関数は、複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた所定のクエリーに対する類似性スコアを計算するために図３から１１に示されている関数の１つ又はそれよりも多くの組合せを利用することができることを認識する。

【００７６】

本発明を好ましい実施形態に関連して説明して例証したが、当業者には明らかなように、本発明の精神及び範囲から逸脱することなく多くの変形及び変更を行うことができ、本発明は、従って、そのような変形及び変更が本発明の範囲に含まれるように意図しているので、上述の方法又は構成の厳密な詳細に制限されないものとする。

【図面の簡単な説明】

【００７７】

【図１】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系の組合せに従って書かれた１つ又はそれよりも多くの関連するクエリーを識別するためのシステムを示すブロック図である。

【図２】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系の組合せに従って書かれた１つ又はそれよりも多くの関連するクエリーを選択する方法の一実施形態を示す流れ図である。

【図３】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーの間の文字編集距離を計算する方法の一実施形態を示す流れ図である。

【図４】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーの間の文字編集距離を計算するための別の実施形態を示す流れ図である。

【図５】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーの間の語編集距離を計算する方法の一実施

形態を示す流れ図である。
【図６】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーに現れる数字の差を識別する方法の一実施形態を示す流れ図である。
【図７】本発明の一実施形態による書記体系の１つのみの文字を考慮に入れて複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーの間の文字編集距離を計算する方法の一実施形態を示す流れ図である。
【図８】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーの接頭辞に重なった文字の数を識別する方法の一実施形態を示す流れ図である。
【図９】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーが非ローマ字文字を有するか否かを識別する方法の一実施形態を示す流れ図である。
【図１０】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーの間の文字編集距離を両方のクエリーが所定の書記体系に変換された後に計算する方法の一実施形態を示す流れ図である。
【図１１】本発明の一実施形態による複数の書記体系を有する言語の１つ又はそれよりも多くの書記体系に従って書かれた２つのクエリーのクエリー及び語句置換確率を計算する方法の一実施形態を示す流れ図である。
【符号の説明】
【００７８】
１００　検索プロバイダ
１０７　検索エンジン
１０８　データ記憶装置
１２２　ネットワーク
１２４ａ、１２４ｂ、１２４ｃ　クライアントデバイス

【図１】

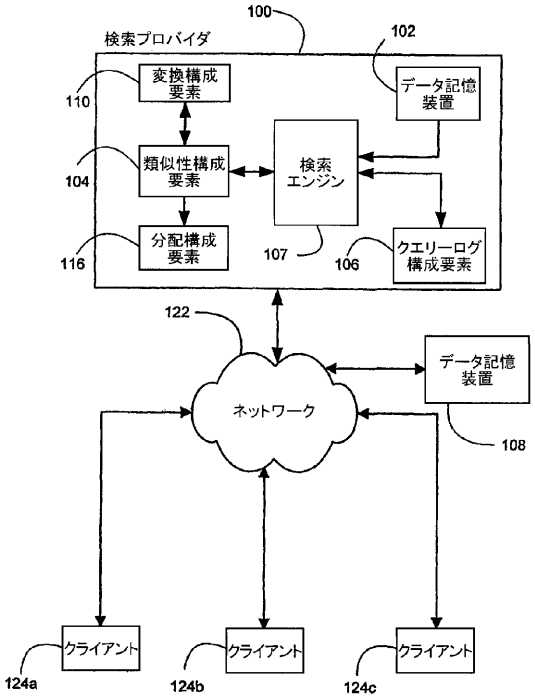



FIG. 1

【図２】

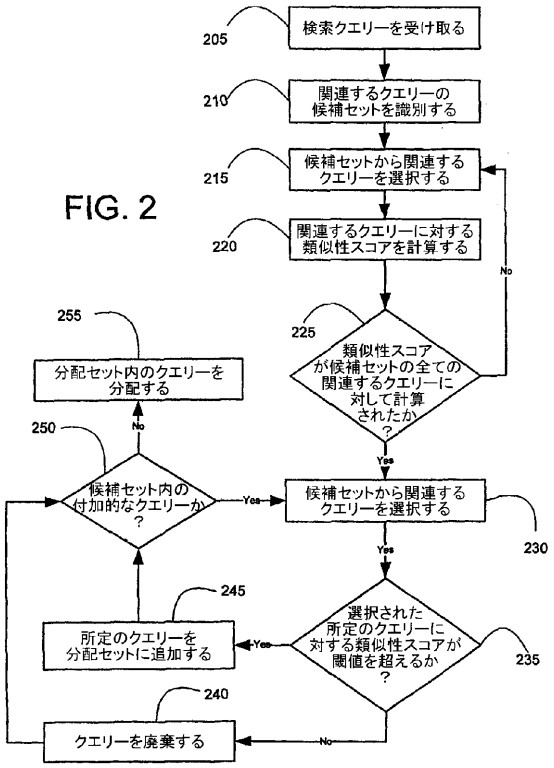



FIG. 2

【図３】

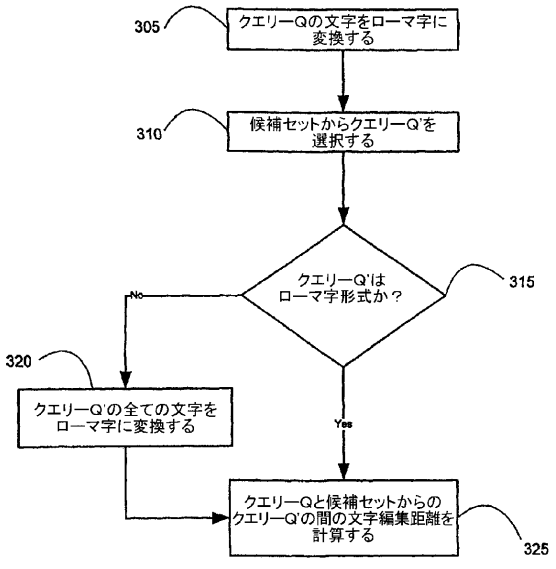



FIG. 3

【図４】

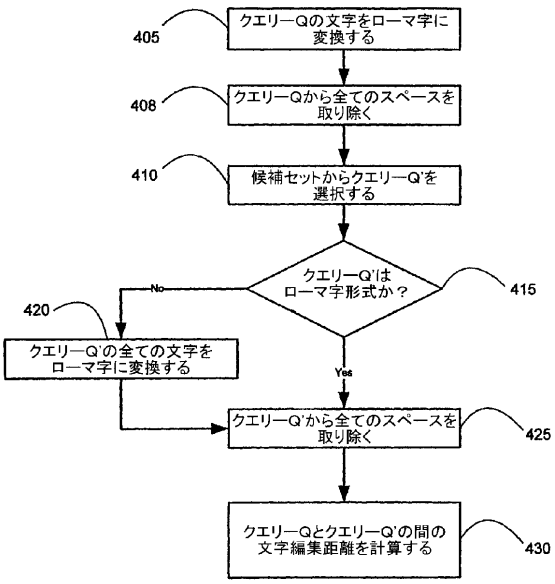



FIG. 4

【図５】

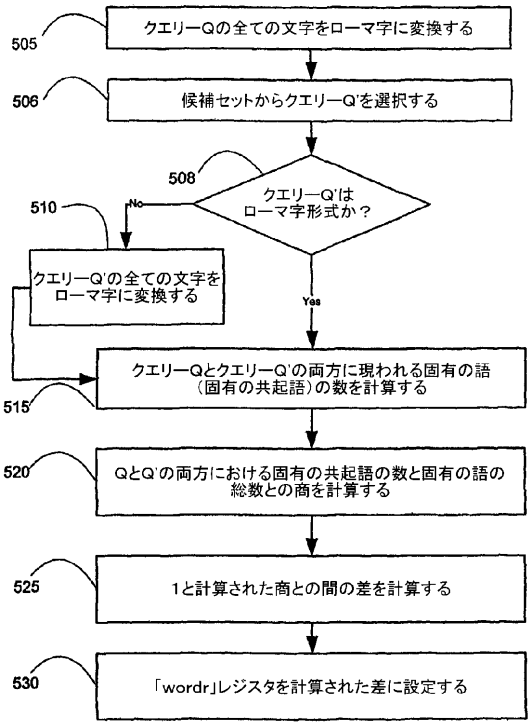



FIG. 5

【図６】

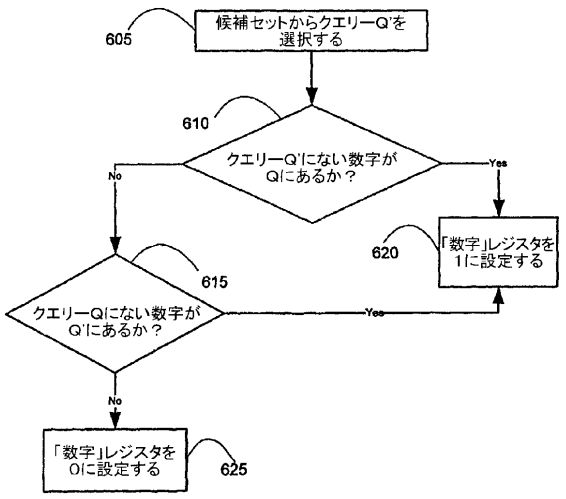



FIG. 6

【図７】

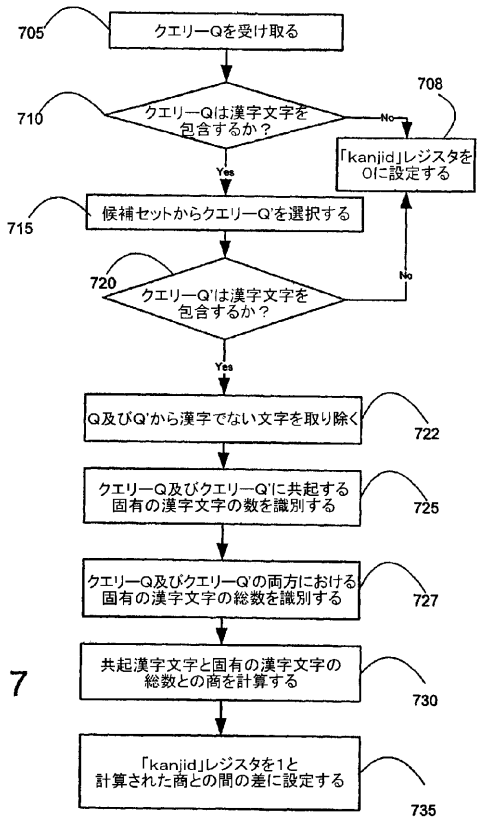



FIG. 7

【図８】

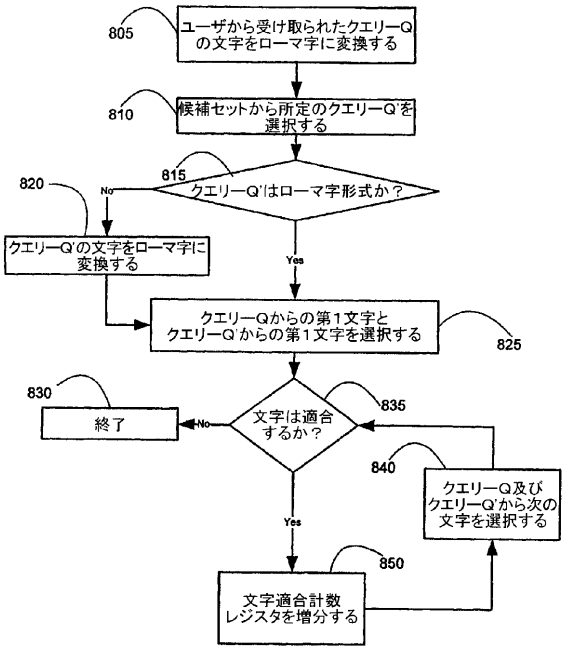



FIG. 8

【図９】

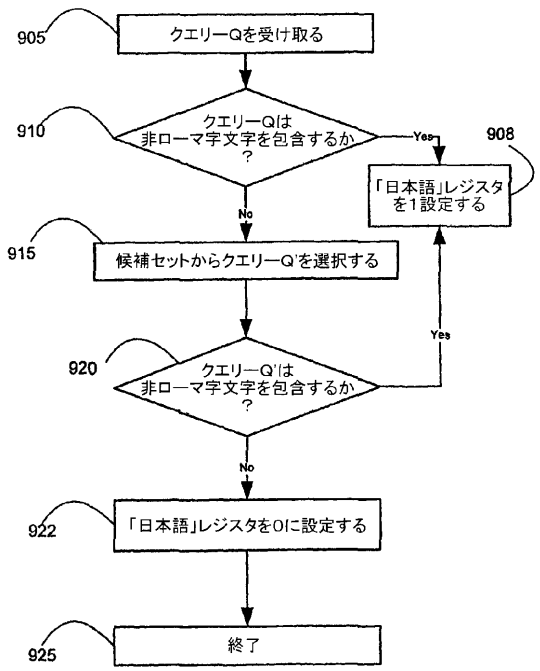



FIG. 9

【図１０】

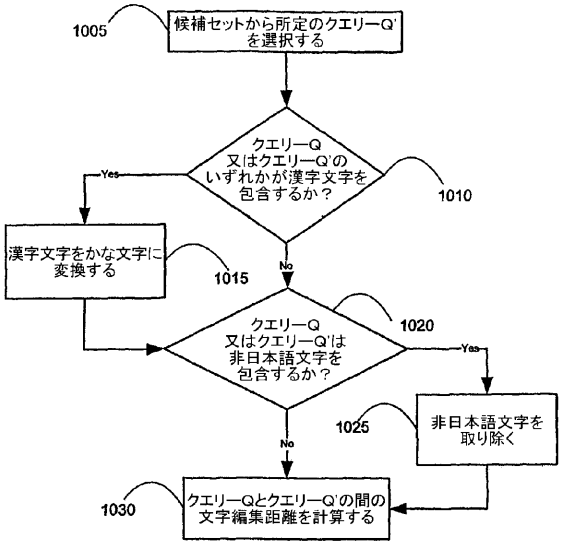



FIG. 10

【図１１】

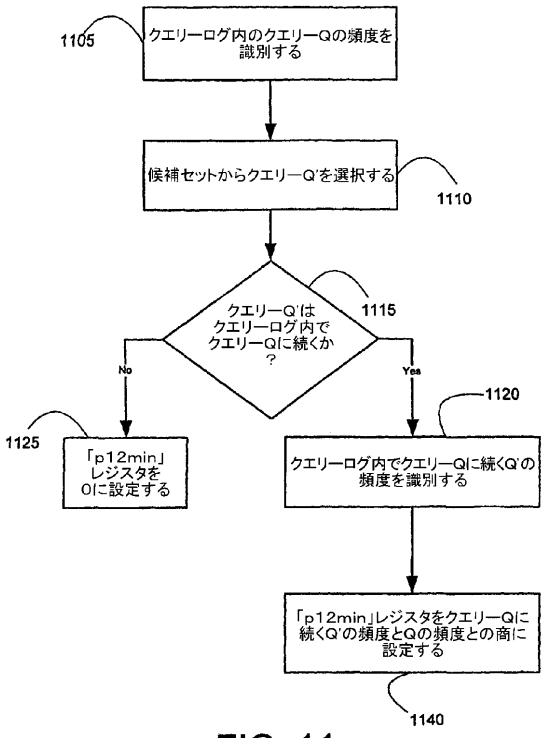



FIG. 11

【国際調査報告】

**INTERNATIONAL SEARCH REPORT**

International application No.
**PCT/US2007/062876**

**A. CLASSIFICATION OF SUBJECT MATTER**

***G06F 17/30(2006.01)i, G06F 17/28(2006.01)i***

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)
IPC 8: G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean Utility models and applications for Utility models since 1975
Japanese Utility models and applications for Utility models since 1975

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKIPASS(KIPO internal), IEEExplore, Google, "Keyword: query, language, identify, similarity, and the similar terms"

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y<br>A | US 2006/0031207 A1 (ANNA BJARNESTAM ET AL.) 09 February 2006<br>see figures 1, 2A, 4; paragraphs 3-4, 17-29; claims 1, 8, 15; | 1-11, 22-25<br>12-21, 26-38 |
| Y<br>A | WO 99/45487 A1 (AMAZON.COM, INC. ET AL.) 10 September 1999<br>see figures 1-2; page 1, line 14-page 3, line 9; claims 1, 3-6; | 1-11, 22-25<br>12-21, 26-38 |
| A | US 06947930 B2 (PETER G. ANICK ET AL.) 20 September 2005<br>see figures 5-6; columns 14-17; claims 1-3, 5-6; | 1-38 |
| A | US 06493709 B1 (ALEXANDER AIKEN) 10 December 2002<br>see figures 1a, 1b, 4a; columns 4-8; claims 2, 8, 9; | 1-38 |
| A | US 05778361 A (TSUTOMU NANJO ET AL.) 07 July 1998<br>see figures 3-9; columns 2-4; claims 1, 6-9; | 1-38 |

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 14 AUGUST 2007 (14.08.2007) | **14 AUGUST 2007 (14.08.2007)** |

Name and mailing address of the ISA/KR
Korean Intellectual Property Office
920 Dunsan-dong, Seo-gu, Daejeon 302-701,
Republic of Korea
Facsimile No. 82-42-472-7140

Authorized officer
SONG, Byoung Jun
Telephone No. 82-42-481-5677

Form PCT/ISA/210 (second sheet) (April 2007)

**INTERNATIONAL SEARCH REPORT**
Information on patent family members

International application No.
**PCT/US2007/062876**

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| US2006031207A1 | 09.02.2006 | JP2006004427A2 | 05.01.2006 |
| WO9945487A1 | 10.09.1999 | AT243869E | 15.07.2003 |
| | | AU1929099A1 | 20.09.1999 |
| | | AU199919290A1 | 20.09.1999 |
| | | AU199919290B2 | 20.09.1999 |
| | | AU757550B2 | 27.02.2003 |
| | | CA2320293AA | 10.09.1999 |
| | | CA2320293C | 03.08.2004 |
| | | CA2320293A1 | 10.09.1999 |
| | | CA2320293C | 10.09.1999 |
| | | DE69815898C0 | 31.07.2003 |
| | | DE69815898T2 | 18.12.2003 |
| | | EP01060449A1 | 20.12.2000 |
| | | EP01060449B1 | 25.06.2003 |
| | | EP1060449A1 | 20.12.2000 |
| | | EP1060449B1 | 25.06.2003 |
| | | EP1060449A1 | 20.12.2000 |
| | | JP14506256 | 26.02.2002 |
| | | JP2002506256T2 | 26.02.2002 |
| | | NZ506229A | 28.02.2003 |
| | | US06185558 | 06.02.2001 |
| | | US2002049752A1 | 25.04.2002 |
| | | US2002049752AA | 25.04.2002 |
| | | US2005177569A1 | 11.08.2005 |
| | | US2005177569AA | 11.08.2005 |
| | | US2006053065AA | 09.03.2006 |
| | | US2007083507AA | 12.04.2007 |
| | | US6185558B1 | 06.02.2001 |
| | | US6185558BA | 06.02.2001 |
| | | US7050992BA | 23.05.2006 |
| | | US7124129BB | 17.10.2006 |
| | | WO9945487A1 | 10.09.1999 |
| US06947930 B2 | 20.09.2005 | EP01606704A2 | 21.12.2005 |
| | | EP1606704A2 | 21.12.2005 |
| | | EP1606704A4 | 26.07.2006 |
| | | JP18523344 | 12.10.2006 |
| | | JP2006523344T2 | 12.10.2006 |
| | | KR1020060002831 | 09.01.2006 |
| | | KR2006002831A | 09.01.2006 |
| | | US20040186827A1 | 23.09.2004 |
| | | US2004186827A1 | 23.09.2004 |
| | | US2004186827AA | 23.09.2004 |
| | | US2006010126AA | 12.01.2006 |
| | | US6947930BB | 20.09.2005 |
| | | WO2004086192A2 | 07.10.2004 |
| | | WO2004086192A3 | 17.02.2005 |

Form PCT/ISA/210 (patent family annex) (April 2007)

**INTERNATIONAL SEARCH REPORT**
Information on patent family members

International application No.
**PCT/US2007/062876**

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| US06493709 B1 | 10.12.2002 | AU199951366A1 | 21.02.2000 |
| | | AU5136699A1 | 21.02.2000 |
| | | US6493709BA | 10.12.2002 |
| | | W00007094A2 | 10.02.2000 |
| | | W0200007094A2 | 10.02.2000 |
| | | W020000709402 | 13.07.2000 |
| | | W0200007094A3 | 10.10.2002 |
| US05778361 A | 07.07.1998 | US5778361A | 07.07.1998 |

Form PCT/ISA/210 (patent family annex) (April 2007)

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IS,IT,LT,LU,LV,MC,NL,PL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BW,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IS,JP,KE,KG,KM,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LT,LU,LV,LY,MA,MD,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RS,RU,SC,SD,SE,SG,SK,SL,SM,SV,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,ZA,ZM,ZW

(72)発明者 ジョーンズ ロージー
アメリカ合衆国 カリフォルニア州 ９１１０７ パサディナ サウス オーク アベニュー ８２
(72)発明者 バーツ ケヴィン
アメリカ合衆国 カリフォルニア州 ９１１０１ パサディナ サウス マディソン ４６５
(72)発明者 レイ ベンジャミン
アメリカ合衆国 カリフォルニア州 ９０４０５ サンタ モニカ コープランド コート ６５５

Ｆターム(参考) 5B075 PP25 PP26