(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5743938号
(P5743938)
(45)発行日　平成27年7月1日(2015.7.1)　(24)登録日　平成27年5月15日(2015.5.15)

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

請求項の数 15　(全 16 頁)

(21)出願番号　特願2012-69750(P2012-69750)
(22)出願日　平成24年3月26日(2012.3.26)
(65)公開番号　特開2013-200795(P2013-200795A)
(43)公開日　平成25年10月3日(2013.10.3)
審査請求日　平成26年6月4日(2014.6.4)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100091096
弁理士　平木　祐輔
(74)代理人　100105463
弁理士　関谷　三男
(74)代理人　100102576
弁理士　渡辺　敏章
(72)発明者　今一　修
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所　中央研究所内

審査官　野崎　大進

最終頁に続く

(54)【発明の名称】連想検索システム、連想検索サーバ及びプログラム

(57)【特許請求の範囲】
【請求項１】
検索要求文書を入力する入力手段と、検索された検索結果を表示する検索結果表示手段とを少なくとも有する検索クライアントと、
複数の文書を格納した文書データベースと、
受信した検索要求文書に対する関連度が高い文書を前記文書データベースから検索する検索手段と、与えられた文書群から特徴単語群を抽出すると共に、抽出された特徴単語群を各単語の重要度とその出現位置情報とに基づいて１つ又は複数の特徴単語群に分類する特徴単語抽出手段とを有し、複数の特徴単語群が抽出された場合、分類後の特徴単語群のそれぞれについて関連度が高い文書を前記文書データベースから検索する連想検索サーバと
を有する連想検索システム。
【請求項２】
請求項１に記載の連想検索システムにおいて、
特徴単語群の分類数は、ユーザがインタフェースを通じて任意に入力する
ことを特徴とする連想検索システム。
【請求項３】
請求項２に記載の連想検索システムにおいて、
前記インタフェースは、特徴単語群の分類数の変更を指示するボタンを有する
ことを特徴とする連想検索システム。

【請求項４】
請求項１に記載の連想検索システムにおいて、
特徴単語群の分類数は、分類スコアと閾値の比較により自動設定される
ことを特徴とする連想検索システム。
【請求項５】
請求項１に記載の連想検索システムにおいて、
前記特徴単語抽出手段は、複数回出現する単語が被覆する範囲を求め、被覆度の少ない箇所で特徴単語群を分類する
ことを特徴とする連想検索システム。
【請求項６】
請求項１に記載の連想検索システムにおいて、
前記特徴単語抽出手段は、複数回出現する単語の重心位置を求め、その重心位置を中心に特徴単語群を分類する
ことを特徴とする連想検索システム。
【請求項７】
請求項１に記載の連想検索システムにおいて、
前記検索クライアントは、検索された文書群の特徴単語を表示する特徴単語表示手段を有する
ことを特徴とする連想検索システム。
【請求項８】
検索クライアントから入力された検索要求文書に類似する文書を複数の文書を格納した文書データベースから検索する連想検索サーバにおいて、
受信した検索要求文書に対する関連度が高い文書を前記文書データベースから検索する検索手段と、
与えられた文書群から特徴単語群を抽出すると共に、抽出された特徴単語群を各単語の重要度とその出現位置情報とに基づいて１つ又は複数の特徴単語群に分類する特徴単語抽出手段と
を有し、複数の特徴単語群が抽出された場合、分類後の特徴単語群のそれぞれについて関連度が高い文書を前記文書データベースから検索する連想検索サーバ。
【請求項９】
請求項８に記載の連想検索サーバにおいて、
前記特徴単語抽出手段は、ユーザがインタフェースを通じて任意に指定した分類数に基づいて特徴単語群を分類する
ことを特徴とする連想検索サーバ。
【請求項１０】
請求項９に記載の連想検索サーバにおいて、
前記インタフェースは、特徴単語群の分類数の変更を指示するボタンを有する
ことを特徴とする連想検索サーバ。
【請求項１１】
請求項８に記載の連想検索サーバにおいて、
前記特徴単語抽出手段は、分類スコアと閾値の比較により、特徴単語群の分類数を自動設定する。
ことを特徴とする連想検索サーバ。
【請求項１２】
検索クライアントから入力された検索要求文書に類似する文書を複数の文書を格納した文書データベースから検索する連想検索サーバとして機能するコンピュータに、
受信した検索要求文書に対する関連度が高い文書を前記文書データベースから検索する第１の処理と、
与えられた文書群から特徴単語群を抽出すると共に、抽出された特徴単語群を各単語の重要度とその出現位置情報とに基づいて１つ又は複数の特徴単語群に分類する第２の処理

と、
複数の特徴単語群が抽出された場合、分類後の特徴単語群のそれぞれについて関連度が高い文書を前記文書データベースから検索する第３の処理と
を実行させるプログラム。
【請求項１３】
請求項１２に記載のプログラムにおいて、
前記第２の処理は、ユーザがインタフェースを通じて任意に指定した分類数に基づいて特徴単語群を分類する
ことを特徴とするプログラム。
【請求項１４】
請求項１３に記載のプログラムにおいて、
前記インタフェースは、特徴単語群の分類数の変更を指示するボタンを有する
ことを特徴とするプログラム。
【請求項１５】
請求項１２に記載のプログラムにおいて、
前記第２の処理は、分類スコアと閾値の比較により、特徴単語群の分類数を自動設定する
ことを特徴とするプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、検索要求として与えられた文書に関連する文書を検索する連想検索システムに関し、特に、与えられた文書中の特徴単語の出現位置の情報を用いる連想検索システム、連想検索サーバ及びそれらを実現するプログラムに関する。
【背景技術】
【０００２】
コンピュータやインターネットの普及に伴い、文書情報の電子化が急速に進んでいる。一方、入手可能な情報の増加に伴い、それらの中から必要な情報を探し出すことが重要な課題となってきている。また、複数の文書データベース間での文書群の関連性を調べたいという要求も高まっている。例えば、興味のある新聞記事に対し、それらに関連する百科事典の項目を検索したいという要求は多い。
【０００３】
現在実用化されているキーワード検索技術の場合、複数の文書データベースを切り替えて検索することは可能であるが、ある文書データベースに含まれる文書群に対し、それに関連する文書群を、同一の文書データベース、あるいは、別の文書データベースから検索すること（文書連想検索と呼ばれる検索方式）は不可能である。
【０００４】
同一の文書データベースに限れば、文書間の類似度を予め計算しておくことにより、文書群を検索入力とした文書連想検索を実現することはできる。しかし、複数の文書データベース間での文書連想検索を実現しようとすると、予め計算すべき文書間の関連度の組み合わせ数が、文書データベース数の増加に伴って爆発的に増加する。このため、文書間の類似度を予め計算する方法による文書連想検索の現実は不可能である。
【０００５】
これに対し、特許文献１には、利用者が指定した文書データベース中の任意の文書群に対して、その文書群に関連する文書群を任意の文書データベースから効率よく検索するための方法が開示されている。
【０００６】
特許文献１に開示の方法は、文書群として入力された検索入力内の特徴的な単語群（特徴単語群）のみを使用し、高速な文書連想検索を実現する。この方法を用いれば、利用者は、複数の異なる種類の文書データベースを切り替えながら、文書群の関連性を調べるこ

とができ、高精度かつ効率的に文書を検索することができる。また、この方法は、検索結果として得られた文書群に出現する特徴単語群を抽出し、それらを検索結果の概観（要約）として利用者に提示することにより、利用者による検索結果の可否の判断を支援する技術も提供する。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】特開２０００－１５５７５８号公報

【発明の概要】

【発明が解決しようとする課題】

【０００８】

一般に、単語に基づく文書検索では、文書中に出現する単語によって文書のインデックス付けを行ない、文書検索を実現する。特許文献１の場合も同様であり、文書から特徴単語群を抽出する際には、文書に含まれる単語の統計的尺度（tf*idf法などが代表的）を用いて重要度を計算し、重要度の高い順に単語を抽出し、連想検索を実現する。

【０００９】

しかし、従来の連想検索では、特徴単語を抽出する対象は文書全体である。このため、文書に複数の話題が含まれている場合には、複数の話題の特徴単語が混在した状態のまま単語が抽出される。つまり、複数の話題を総合的に判断して類似した文書が検索される。このため、利用者が望んだ結果が必ずしも得られるとは限らない。

【００１０】

この技術課題を鑑み、本発明は、検索入力となる文書群中に含まれる話題ごとに類似する文書を検索できる連想検索システムを提供する。

【課題を解決するための手段】

【００１１】

このために、本発明においては、連想検索における検索入力文書から特徴単語群を抽出する際に、各単語の重要度だけでなく、その単語の検索入力文書中での位置情報も付加して抽出処理を実行する。次に、抽出した特徴単語群を、各単語の重要度と出現位置に基づいて分類する。特徴単語群の分類数は、検索入力文書中での特徴単語の重要度と距離に応じて分類する際の分類スコアに閾値を設定して自動的に設定してもよいし、利用者がユーザインタフェース上で分類数を任意に設定してもよい。最後に、分類結果として得られた特徴単語群のそれぞれを検索入力として検索を実行する。

【発明の効果】

【００１２】

本発明によれば、複数の話題が含まれている文書群を検索入力とする場合でも、文書群全体として類似した文書ではなく、分類された特徴単語群（文書中に含まれる話題に相当）毎に類似した文書を連想検索結果として得ることができる。これにより、利用者の希望により近い結果を提示することができる。前述した以外の課題、構成及び効果は、以下の実施の形態の説明により明らかにされる。

【図面の簡単な説明】

【００１３】

【図１】連想検索システムの概念的な構成例を示す図。

【図２】連想検索スレーブサーバの構成例を示す図。

【図３】検索手段の構成例を示す図。

【図４】特徴単語抽出手段の構成例を示す図。

【図５】文書中における特徴単語の分布例を示す図。

【図６】特徴単語の抽出及び分類方法の例を示す図。

【図７】特徴単語の抽出及び分類方法の例を示す図。

【図８】検索クライアントにおける初期画面の例を示す図。

【図９】検索クライアントにおける検索結果の表示例を示す図。

【図１０】検索クライアントにおける検索結果の表示例を示す図。
【図１１】検索クライアントにおける検索結果の表示例を示す図。
【図１２】特徴単語群分類結果の確認画面の表示例を示す図。
【図１３】インデックス付けの一例を示す図。
【図１４】検索クライアント、連想検索マスタサーバ、連想検索スレーブサーバ間におけるデータ及び処理の流れを示す図。
【図１５】検索クライアント、連想検索マスタサーバ、連想検索スレーブサーバ間におけるデータ及び処理の流れを示す図。
【図１６】検索クライアント、連想検索マスタサーバ、連想検索スレーブサーバ間におけるデータ及び処理の流れを示す図。
【発明を実施するための形態】
【００１４】
以下、図面に基づいて、本発明の実施の形態を説明する。なお、本発明の実施の態様は、後述する形態例に限定されるものではなく、その技術思想の範囲において、種々の変形が可能である。
【００１５】
図１は、形態例に係る連想検索システムの概略構成を示している。このシステムは、利用者による検索要求の入力及び検索結果の表示に使用される検索クライアント２０と、文書データベースを検索する連想検索スレーブサーバ４０、５０、６０と、連想検索クライアント２０と連想検索スレーブサーバ４０、５０、６０を仲介する連想検索マスタサーバ３０と、これらを接続する通信ネットワーク１０とで構成される。
【００１６】
図１の例では、文書データベースを検索する連想検索スレーブサーバが通信ネットワーク１０に３台接続されている場合を表しているが、通信ネットワーク１０に接続される連想検索スレーブサーバの数は任意でよい。検索クライアント２０の数も任意である。
【００１７】
また、図１の例では、検索クライアント１０と、連想検索マスタサーバ３０と、連想検索スレーブサーバ４０、５０、６０とを通信ネットワーク１０を介して接続しているが、これらのうちの幾つかを、あるいは、全てを同一の計算機上に構成してもよい。
【００１８】
図２に、連想検索スレーブサーバ４０の構成例を示す。他の連想検索スレーブサーバ５０、６０の構成も、連想検索スレーブサーバ４０と同じである。連想検索スレーブサーバ４０は、メモリ装置４９１、演算処理装置４９２、インタフェース装置４９３、補助記憶装置４９４、入力装置４９５、出力装置４９６を有し、それぞれがバス４９０を介して相互に接続されている。
【００１９】
メモリ装置４９１は、補助記憶装置４９４からプログラムを読み出して記憶するＲＡＭ（Random Access Memory）等の記憶装置である。メモリ装置４９１には、検索手段４１０と特徴単語抽出手段４２０に対応するプログラム、その実行に必要な検索インデックス４３０と文書データベース４４０に対応するファイルやデータ等が記憶される。
【００２０】
演算処理装置４９２は、メモリ装置４９１に格納されたプログラムを実行するＣＰＵ（Central Processing Unit）等の演算処理装置である。インタフェース装置４９３は、外部ネットワーク等に接続するためのインタフェース装置である。補助記憶装置４９４は、検索手段４１０と特徴単語抽出手段４２０に対応するプログラム、検索インデックス４３０と文書データベース４４０に対応するファイルやデータ等を記憶するＨＤＤ（Hard Disk Drive）等の記憶装置である。入力装置４９５は、ユーザインタフェースを提供する装置（例えば、キーボード、マウス）である。出力装置４９６は、ユーザインタフェースを提供する出力装置（例えば、ディスプレイ装置）である。
【００２１】

図２は、連想検索スレーブサーバの構成を示す図であるが、検索クライアント２０と連想検索マスタサーバ３０の構成も、補助記憶装置に記憶されるプログラムやデータの違いを除き、同様に構成される。

【００２２】

図３に、連想検索スレーブサーバ４０が備える検索手段４１０の機能ブロック構成を示す。プログラムとしての検索手段４１０は、単語頻度取得手段４１１、位置情報取得手段４１２、関連度計算手段４１３、近似性計算手段４１４、スコア計算手段４１５の各機能により構成される。これらの検索手段４１０を構成する各手段もプログラム処理を通じて提供される。

【００２３】

連想検索スレーブサーバ４０は、連想検索マスタサーバ３０が備える検索要求発行手段３２０から送られてきた検索要求に対し、関連度の高い文書群を文書データベース４４０から検索し、その検索結果を関連度のスコア付きで連想検索マスタサーバ３０に返す。ここでの検索は、例えば公知のキーワード検索手法により実現することができる。

【００２４】

キーワード検索手法では、検索処理の効率を上げるために、文書データベースに含まれる文書を単語に分割し（日本語の文書に対しては形態素解析を実行し、英語の文書に対してはステミング処理を実行する）、どの文書にどの単語が含まれているかを示す検索インデックスを事前に作成する。後述する本実施例の検索方法のように、検索時に位置情報も用いる場合には、各単語の出現位置もインデックスに格納しておく。検索実行時には、事前に作成された検索インデックスを用いることで、検索処理を高速に実行することができる。

【００２５】

図１の場合には、連想検索スレーブサーバ４０、５０、６０が有する文書データベース４４０、５４０、６４０のそれぞれについて、検索インデックス４３０、５３０、６３０を事前に作成し、検索処理に利用する。

【００２６】

検索要求と検索対象文書間の関連度の計算は、以下の手順で実行される。まず、検索手段４１０が、連想検索マスタサーバ３０の検索要求発行手段３２０から送信された検索要求を受信する。検索手段４１０は、受信した検索要求に含まれる単語群を含む文書を検索する。単語頻度取得手段４１１は、検索結果として得られた文書のそれぞれについて、各文書に含まれる単語群のうち検索要求に含まれる単語群の頻度情報を取得する。次に、関連度計算手段４１３は検索要求とその文書の関連度を計算する。関連度の計算方法は任意でよい。例えば公知の技術であるtf*idf法により単語の重要度を計算し、その総和を関連度とする。単語の近接性を検索スコアに反映する場合には、位置情報取得手段４１２が、各文書に含まれる単語群のうち検索要求に含まれる単語群の出現位置情報を取得し、近接性計算手段４１４が近接スコアを計算する。近接スコアの計算方法は任意でよい。例えば、検索要求に含まれる単語群がどれくらい密集して出現しているかを計算し、その計算結果を近接スコアとする。スコア計算手段４１５は、関連度計算手段４１３と近接性計算手段４１４のそれぞれから得られたスコアを統合し、統合後のスコアを関連度として文書に付与する。

【００２７】

図４に、連想検索スレーブサーバ４０が備える特徴単語抽出手段４２０の機能ブロック構成を示す。プログラムとしての特徴単語抽出手段４２０は、単語頻度取得手段４２１、位置情報取得手段４２２、重要度計算手段４２３、近接性クラスタリング手段４２４、単語追加手段４２５の各機能により構成される。これらの特徴単語抽出手段４２０を構成する各手段もプログラム処理を通じて提供される。

【００２８】

特徴単語抽出手段４２０は、連想検索マスタサーバ３０が備える特徴単語要求手段３３０から送られてきた文書群に対する特徴単語を、文書データベース４４０から抽出する。

特徴単語抽出手段４２０は、特徴単語の高速抽出を実現するために、検索手段４１０と同様、検索インデックス４３０を利用する。すなわち、特徴単語抽出手段４２０は、ある文書にどの単語が含まれているかを、検索インデックス４３０を参照して調べる。

【００２９】

特徴単語の抽出は、以下の手順で実行される。まず、特徴単語抽出手段４２０が、連想検索マスタサーバ３０の特徴単語要求手段３３０から送信された文書群を受信する。単語頻度取得手段４２１は、受信した文書群に含まれる各単語の頻度情報を取得する。取得された頻度情報に基づいて、重要度計算手段４２３は、各単語の重要度を計算する。重要度の計算方法は任意でよい。例えば公知の技術であるtf*idf法により単語の重要度を計算する。位置情報を用いない連想検索の場合、特徴単語抽出手段４２０は、高い重要度が付された単語から順番に特徴単語として連想検索マスタサーバ３０に返す。

【００３０】

本実施の形態では、位置情報取得手段４２２が、重要度付きの各単語について出現位置情報を取得する。さらに、近接性クラスタリング手段４２４が、重要度と位置情報とに基づいて検索された単語群を分類する。さらに、単語追加手段４２５が、分類結果のそれぞれに含まれる単語群に近接する単語を追加する。特徴単語抽出手段４２０は、このようにして得られた特徴単語群の集合を連想検索マスタサーバ３０に返す。単語追加手段４２５の使用は任意でよい。

【００３１】

次に、近接性クラスタリング手段４２４の動作を図５、図６、図７を用いて説明する。図５は、同じ単語が含まれる二つの文書１、文書２を例示している。文書１では、各単語（term1～term6）が文書全体に分散して分布しているのに対し、文書２では、term1～term3が文書中の前半に、term4～term6が文書中の後半に集中して分布している。

【００３２】

このような場合でも、従来の連想検索では、特徴単語の出現位置を考慮していないため、文書１を検索入力として連想検索を実行した場合の結果と、文書２を検索入力として連想検索を実行した場合の結果は、同じである。しかし、特徴単語の文書中での分布が偏っている場合、複数の話題について書かれている可能性があるため、特徴単語群を個々の話題に分類することが望ましい。

【００３３】

図６は、文書１から特徴単語群を抽出する場合の例である。この場合、各特徴単語は、文書全体に分散して分布しているため、一つの話題について書かれていると考えられる。従って、位置情報に基づいて特徴単語群を分類しても、分類することができず、一つの特徴単語群となる。

【００３４】

図７は、文書２から特徴単語群を抽出する場合の例である。この場合、各特徴単語は、文書の前半にterm1～term3、文書の後半にterm4～term6が集中して分布しているため、二つの話題について書かれていると考えられる。従って、位置情報に基づいて特徴単語群を分類すると、term1～term3の特徴単語群と、term4～term6の特徴単語群の二つの特徴単語群が抽出される。

【００３５】

近接性クラスタリング手段４２４による特徴単語群の分類には、例えば、単語の出現位置とその重みを用いる階層的クラスタリング手法を適用すればよい。複数回出現する単語については、予め、その重心位置を求めておく。その後、各単語の位置に基づいて、最も近接する単語をまとめあげる。その際、それぞれに単語の重みを考慮して、新しい重心を決定する。この処理を繰り返すことでクラスタリング結果を得る。

【００３６】

あるいは、別の手法として、複数回出現する単語が文書中のどの範囲を被覆するかを求め、文書全体における被覆度の少ない箇所で特徴単語群を分類してもよい。

【００３７】

前述の説明では、近接性クラスタリング手段４２４における特徴単語分類手法として、二つの手法について説明したが、位置情報に基づいて特徴単語群を分類する手法であれば任意のものを用いてもよい。

【００３８】

このようにして得られた特徴単語群を用いて連想検索を実行することにより、文書中に複数の話題が含まれている場合でも、利用者の望んだ検索結果を得ることが可能となる。

【００３９】

図８は、検索クライアント２０が備える検索要求入力手段２１０により提供される画面例を表している。利用者は、検索要求入力エリア２１１に検索要求を入力し、検索指示ボタン２１２をクリックすることにより検索の実行を検索クライアント２０に指示する。

【００４０】

図９は、検索クライアント２０による検索結果の表示例である。検索結果は、検索結果表示手段２２０により表示され、検索結果から抽出された特徴単語群が特徴単語表示手段２３０により表示される。特徴単語表示手段２３０を用いるか否かは任意である。検索結果表示手段２２０は文書群指定手段も兼ねている。文書選択チェックボックス２２１により任意個の文書を選択した状態で、連想検索指示ボタン２１３をクリックすると、選択した文書と関連する文書を検索することができる。特徴単語表示手段２３０は、単語群指定手段も兼ねている。単語選択チェックボックス２３１により任意個の単語を選択した状態で、連想検索指示ボタン２１３をクリックすると、特徴単語からの検索を実行することができる。分類数指定手段２４０は、文書を選択して連想検索を実行する場合に、文書中に含まれる話題を何個に分割するかを指定入力するために用いられる。分類数は、数値として直接指定してもよいし、スライドバーやボタン等を用いて指定してもよい。また、分類数は、分類スコアと閾値との比較を通じて自動的に設定してもよい。分類スコアは、特徴単語の重要度のスコアと近接度合のスコアを統合したスコアとして規定する。分類数を閾値により自動設定する場合には、分類数指定手段２４０を画面に表示しなくてもよい。

【００４１】

図１０は、検索入力として与えられた文書に二つの話題が含まれている場合の検索結果の例である。この場合、検索結果表示手段２２０には、二列に分けて、それぞれの話題に関する検索結果が表示される。左列の記事１～５が話題１に対応し、右列の記事Ａ～Ｅが話題２に対応する。なお、図１０の場合、特徴単語表示手段２３０には、二つの話題の検索結果を統合して、そこから特徴単語群を抽出した結果を表示している。

【００４２】

一方、図１１は、検索入力として与えられた文書に二つの話題が含まれている点は図１０と同じであるが、特徴単語表示手段２３０に、各話題の検索結果ごとに特徴単語群を抽出し、それぞれを二列に表示している。左列の特徴ターム１～５が話題１に対応し、右列の特徴タームＡ～Ｅが話題２に対応する。図９の場合と同様、特徴単語表示手段２３０を用いるか否かは任意である。

【００４３】

図１２は、近接性クラスタリング手段４２４が分類した特徴単語群を確認する画面である。利用者は、この画面を用いて、分類された特徴単語群が適切かどうかを判断し、適切であれば検索指示ボタン２１３をクリックする。適切でなければ、利用者は、分類数指定手段２４０に新たな分類数を指定し、その後、分類数変更指示ボタン２４１をクリックし、再度、分類された特徴単語群を確認する。なお、この画面の使用は任意である。

【００４４】

図１３は、文書データベース４４０、５４０、６４０に含まれる文書から検索インデックス４３０、５３０、６３０を作成した場合の検索インデックスの例である。文書ＩＤの列に個々の文書を識別する識別子、その識別子に該当する文書に含まれる単語の出現位置の情報が格納されている。

【００４５】

次に、実施の形態に係る連想検索システムで実行される処理の流れを、図１４のシーケ

ンス図を用いて説明する。以下では、連想検索スレーブサーバとして連想検索スレーブサーバ４０を用いる場合を説明する。

【００４６】

利用者は、検索クライアント２０が備える検索要求入力手段２１０を用い、検索要求を入力する。入力された検索要求は、検索クライアント２０から連想検索マスタサーバ３０に送信される（Ｔ１１）。

【００４７】

連想検索マスタサーバ３０の検索要求解析手段３１０は検索要求を解析し、連想検索スレーブサーバ４０に送信するための検索要求を作成する。検索要求発行手段３２０により、検索要求が連想検索スレーブサーバ４０に送信される（Ｔ１２）。

【００４８】

連想検索スレーブサーバ４０が備える検索手段４１０は、検索インデックス４３０を用いて文書データベース４４０を検索し、その結果を連想検索マスタサーバ３０に返す（Ｔ１３）。

【００４９】

連想検索マスタサーバ３０の特徴単語要求手段３３０は、得られた検索結果から特徴単語を抽出するために、特徴単語の抽出要求を連想検索スレーブサーバ４０に送信する（Ｔ１４）。

【００５０】

連想検索スレーブサーバ４０が備える特徴単語抽出手段４２０は、検索インデックス４３０を利用して特徴単語群を抽出し、連想検索マスタサーバ３０へ返す（Ｔ１５）。

【００５１】

最後に、検索結果と特徴単語群が連想検索マスタサーバ３０から検索クライアント２０に送信され（Ｔ１６）、検索クライアント２０の検索結果表示手段２２０と特徴単語表示手段２３０によって利用者に提示される。

【００５２】

次に、図１５に示すシーケンス図について説明する。このシーケンス図は、検索結果として得られた文書群から連想検索を実行する場合の処理の流れを示している。

【００５３】

利用者は、検索クライアント２０が備える文書群指定手段２２０を用いて、検索入力となる文書群を選択する。選択された文書群の識別子は連想検索マスタサーバ３０に送信される（Ｔ２１）。

【００５４】

連想検索マスタサーバ３０の特徴単語要求手段３３０は、選択された文書群から特徴単語を抽出するために、特徴単語の抽出要求を連想検索スレーブサーバ４０に送信する（Ｔ２２）。

【００５５】

連想検索スレーブサーバ４０が備える特徴単語抽出手段４２０は、検索インデックス４３０を利用して特徴単語群を抽出し、連想検索マスタサーバ３０へ返す（Ｔ２３）。

【００５６】

連想検索マスタサーバ３０の検索要求発行手段３２０は、得られた特徴単語群を連想検索スレーブサーバに送信する（Ｔ２４）。

【００５７】

連想検索スレーブサーバ４０が備える検索手段４１０は、検索インデックス４３０を用いて文書データベース４４０を検索し、その結果を連想検索マスタサーバ３０に返す（Ｔ２５）。

【００５８】

連想検索マスタサーバ３０の特徴単語要求手段３３０は、得られた検索結果から特徴単語を抽出するために、特徴単語の抽出要求を連想検索スレーブサーバ４０に送信する（Ｔ２６）。

【００５９】
連想検索スレーブサーバ４０が備える特徴単語抽出手段４２０は、検索インデックス４３０を利用して特徴単語群を抽出し、連想検索マスタサーバ３０へ返す（Ｔ２７）。
【００６０】
最後に、検索結果と特徴単語群が連想検索マスタサーバ３０から検索クライアント２０に送信され（Ｔ２８）、検索クライアント２０の検索結果表示手段２２０と特徴単語表示手段２３０によって利用者に提示される。
【００６１】
次に、図１６に示すシーケンス図について説明する。このシーケンス図は、検索結果として得られた文書群から連想検索を実行する場合の処理の流れを示しており、かつ、得られた文書群に二つの話題が含まれている場合を示している。
【００６２】
利用者は、検索クライアント２０が備える文書群指定手段２２０を用いて、検索入力となる文書群を選択する。選択された文書群の識別子は、検索クライアント２０から連想検索マスタサーバ３０に送信される（Ｔ３１）。
【００６３】
連想検索マスタサーバ３０の特徴単語要求手段３３０は、選択された文書群から特徴単語を抽出するために、特徴単語の抽出要求を連想検索スレーブサーバ４０に送信する（Ｔ３２）。
【００６４】
連想検索スレーブサーバ４０が備える特徴単語抽出手段４２０は、検索インデックス４３０を利用して特徴単語群を抽出し、連想検索マスタサーバ３０へ返す（Ｔ３３）。
【００６５】
連想検索マスタサーバ３０の検索要求発行手段３２０は、得られた二つの特徴単語群のうち一つ目の話題に相当する特徴単語群を連想検索スレーブサーバ４０に送信する（Ｔ３４１）。
【００６６】
連想検索スレーブサーバ４０が備える検索手段４１０は、検索インデックス４３０を用いて文書データベース４４０を検索し、その結果を連想検索マスタサーバ３０に返す（Ｔ３５１）。
【００６７】
連想検索マスタサーバ３０の特徴単語要求手段３３０は、得られた検索結果から特徴単語を抽出するために、特徴単語の抽出要求を連想検索スレーブサーバ４０に送信する（Ｔ３６１）。
【００６８】
連想検索スレーブサーバ４０が備える特徴単語抽出手段４２０は、検索インデックス４３０を利用して特徴単語群を抽出し、連想検索マスタサーバ３０へ返す（Ｔ３７１）。
【００６９】
次に、連想検索マスタサーバ３０の検索要求発行手段３２０は、得られた二つの特徴単語群のうち二つ目の話題に相当する特徴単語群を連想検索スレーブサーバ４０に送信する（Ｔ３４２）。
【００７０】
連想検索スレーブサーバ４０が備える検索手段４１０は、検索インデックス４３０を用いて文書データベース４４０を検索し、その結果を連想検索マスタサーバ３０に返す（Ｔ３５２）。
【００７１】
連想検索マスタサーバ３０の特徴単語要求手段３３０は、得られた検索結果から特徴単語を抽出するために、特徴単語の抽出要求を連想検索スレーブサーバ４０に送信する（Ｔ３６２）。
【００７２】

連想検索スレーブサーバ４０が備える特徴単語抽出手段４２０は、検索インデックス４３０を利用して特徴単語群を抽出し、連想検索マスタサーバ３０へ返す（Ｔ３７２）。

【００７３】

最後に、検索結果と特徴単語群が連想検索マスタサーバ３０から検索クライアント２０に送信され（Ｔ２８）、検索クライアント２０の検索結果表示手段２２０と特徴単語表示手段２３０によって利用者に提示される。

【００７４】

話題が三つ以上ある場合には、Ｔ３３の後の検索要求発行手段→Ｔ３４１→検索手段→Ｔ３５１→特徴単語要求手段→Ｔ３６１→特徴単語抽出手段→Ｔ３７１と同様の処理を必要な回数繰り返せばよい。

【００７５】

図１０に示したように二つの話題の検索結果全体から特徴単語を抽出する場合は、図１６のシーケンス図において、Ｔ３５１の後の特徴単語要求手段→Ｔ３６１→特徴単語抽出手段→Ｔ３７１を省略し、Ｔ３５２の後の特徴単語要求手段において、二つの話題の検索結果全体の文書群を連想検索スレーブサーバ４０に送信すればよい。

【００７６】

なお、本発明は上述した形態例に限定されるものでなく、様々な変形例が含まれる。例えば、上述した形態例は、本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、ある形態例の一部を他の形態例の構成に置き換えることが可能であり、また、ある形態例の構成に他の形態例の構成を加えることも可能である。また、各形態例の構成の一部について、他の構成を追加、削除又は置換することも可能である。

【００７７】

また、上述した各構成、機能、処理部、処理手段等は、それらの一部又は全部を、例えば集積回路その他のハードウェアとして実現することも可能である。

【符号の説明】

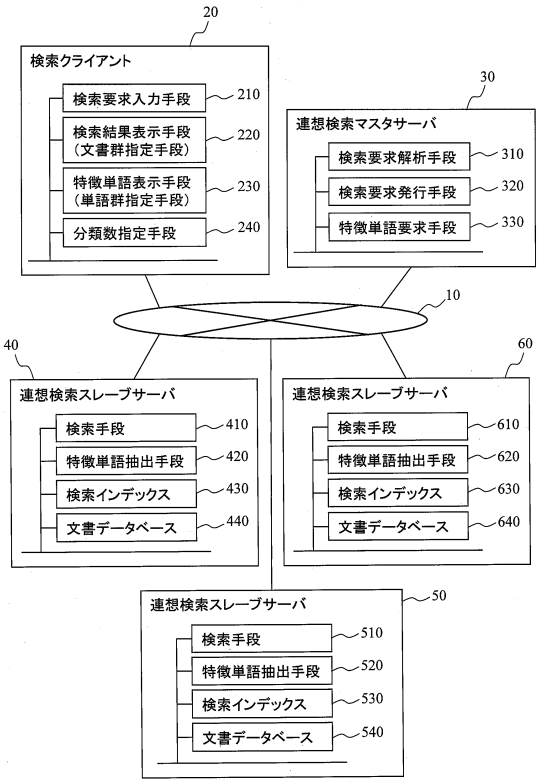
【００７８】

１０：通信ネットワーク
２０：検索クライアント
２１０：検索要求入力手段
２１１：検索要求入力エリア
２１２：検索指示ボタン
２１３：連想検索指示ボタン
２２０：検索結果表示手段（文書群指定手段）
２２１：文書選択チェックボックス
２３０：特徴単語表示手段（単語群指定手段）
２３１：単語選択チェックボックス
２４０：分類数指定手段
２４１：分類数変更指示ボタン
３０：連想検索マスタサーバ
３１０：検索要求解析手段
３２０：検索要求発行手段
３３０：特徴単語要求手段
４０：連想検索スレーブサーバ
４１０：検索手段
４１１：単語頻度取得手段
４１２：位置情報取得手段
４１３：関連度計算手段
４１４：近接性計算手段
４１５：スコア計算手段

４２０：特徴単語抽出手段
４２１：単語頻度取得手段
４２２：位置情報取得手段
４２３：重要度計算手段
４２４：近接性クラスタリング手段
４２５：単語追加手段
４３０：検索インデックス
４４０：文書データベース
４９０：バス
４９１：メモリ装置
４９２：演算処理装置
４９３：インタフェース装置
４９４：補助記憶装置
４９５：入力装置
４９６：出力装置
５０：連想検索スレーブサーバ
５１０：検索手段
５２０：特徴単語抽出手段
５３０：検索インデックス
５４０：文書データベース
５０：連想検索スレーブサーバ
５１０：検索手段
５２０：特徴単語抽出手段
５３０：検索インデックス
５４０：文書データベース

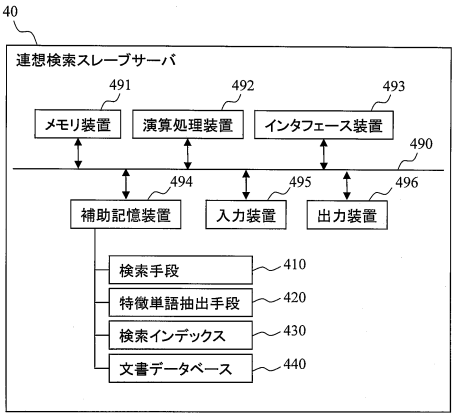
【図１】

図１



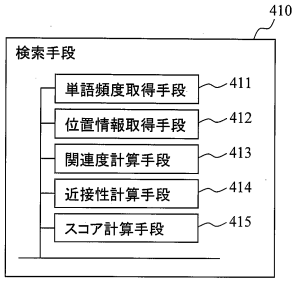
【図２】

図２



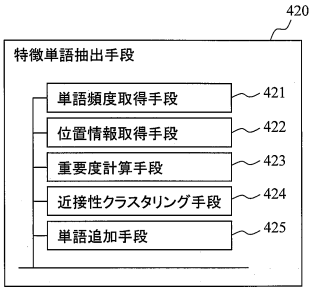
【図３】

図３



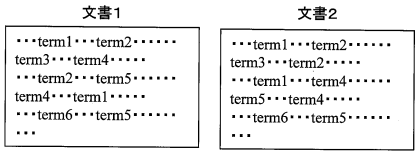
【図４】

図４



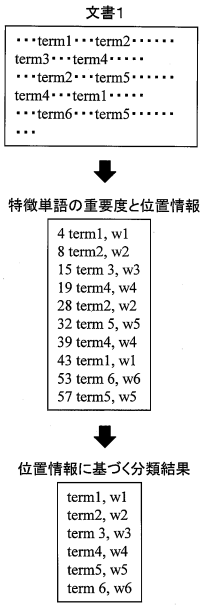
【図５】

図５



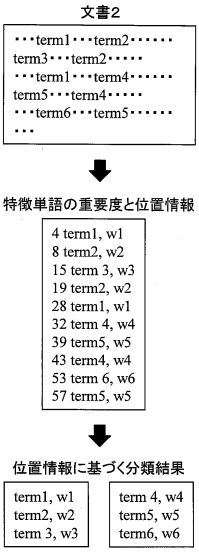
【図６】

図６



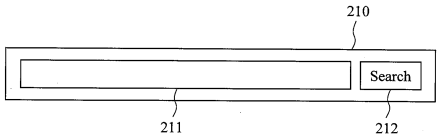
【図７】

図７



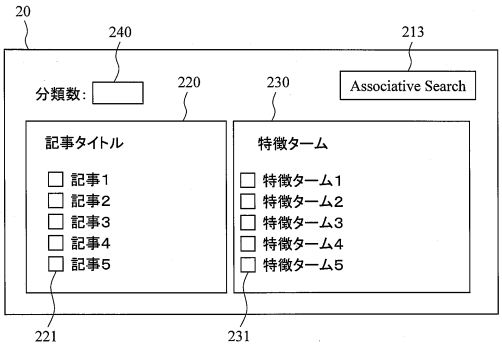
【図８】

図８



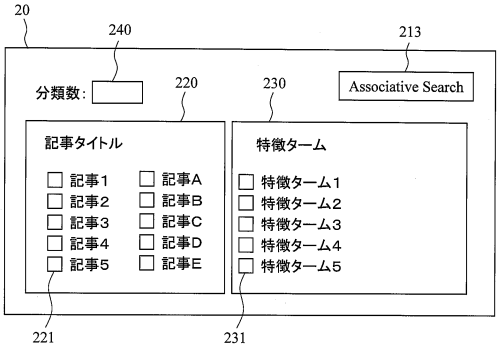
【図９】

図９



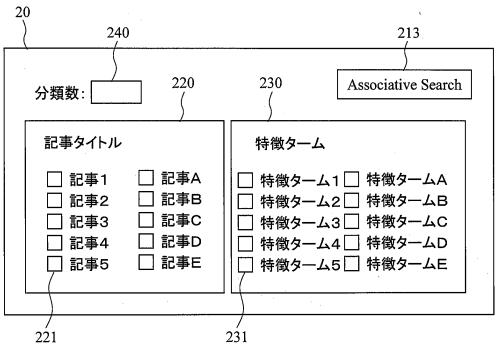
【図１０】

図１０



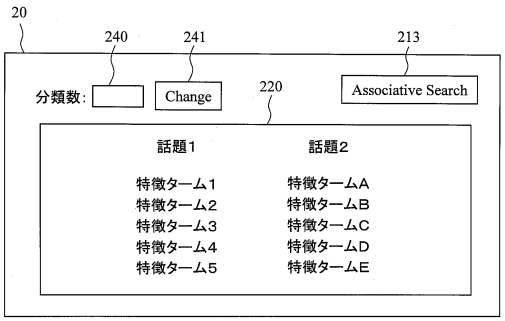
【図１１】

図１１



【図１２】

図１２



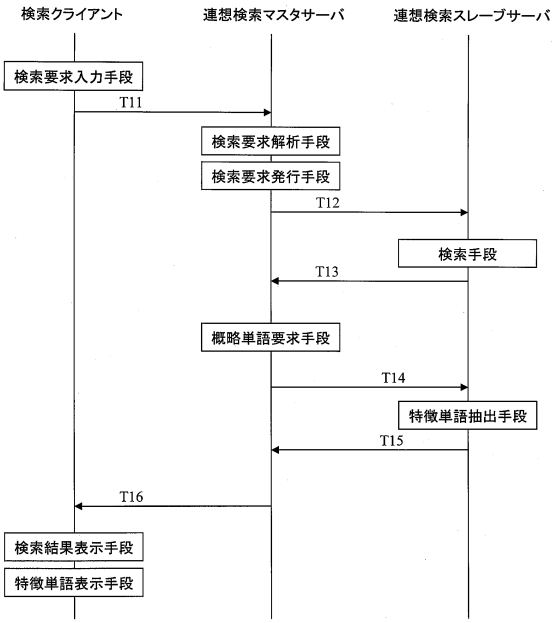
【図１３】

図１３

文書ID	出現位置
123	1,3,7 $term_1$ 2 $term_2$ 4 $term_3$ ・・・ n $term_m$
124	・・・

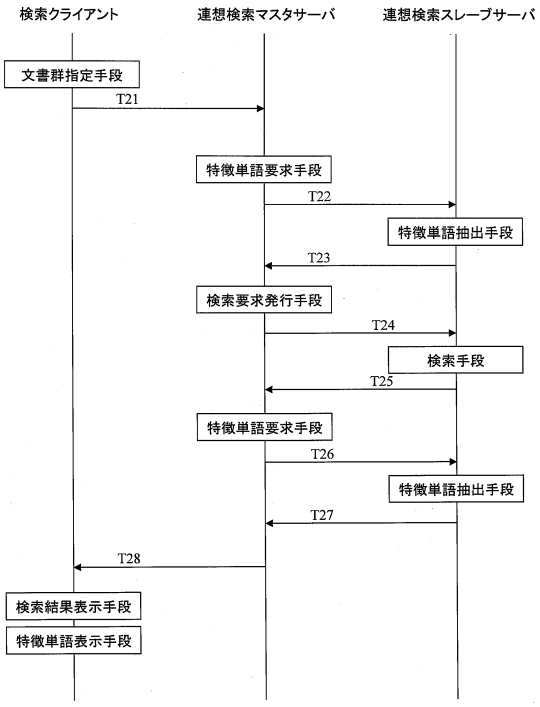
【図１４】

図１４



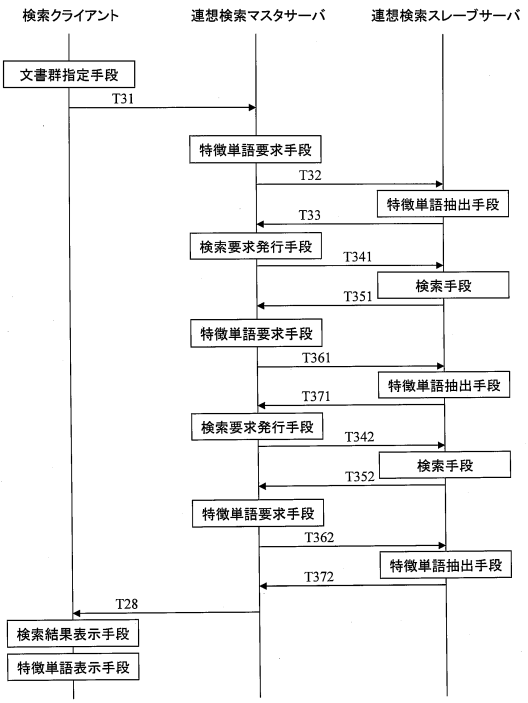
【図１５】

図１５



【図１６】

図１６



フロントページの続き

(56)参考文献 特開２０００－０１０９８６（ＪＰ，Ａ）
特開２００２－２６９１４０（ＪＰ，Ａ）
特開２００７－２６５０３４（ＪＰ，Ａ）
特許第５３８７８７０（ＪＰ，Ｂ２）
欧州特許第００７３０７６５（ＥＰ，Ｂ１）
米国特許出願公開第２００９／０１６９１１０（ＵＳ，Ａ１）

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