(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3760148号
(P3760148)
(45)発行日　平成18年3月29日(2006.3.29)　(24)登録日　平成18年1月13日(2006.1.13)

(51)Int.Cl.　F I
GO6F 17/30 (2006.01)　GO6F 17/30 110C
GO6F 17/30 170A
GO6F 17/30 210A

請求項の数 3　(全 7 頁)

(21)出願番号　特願2002-269885(P2002-269885)
(22)出願日　平成14年9月17日(2002.9.17)
(62)分割の表示　特願平10-328940の分割
原出願日　平成10年11月19日(1998.11.19)
(65)公開番号　特開2003-178095(P2003-178095A)
(43)公開日　平成15年6月27日(2003.6.27)
審査請求日　平成15年11月26日(2003.11.26)

(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　110000350
特許業務法人　日東国際特許事務所
(74)代理人　100068504
弁理士　小川　勝男
(74)代理人　100086656
弁理士　田中　恭助
(72)発明者　岩山　真
埼玉県比企郡鳩山町赤沼２５２０番地　株式会社日立製作所　基礎研究所内
(72)発明者　西岡　真吾
埼玉県比企郡鳩山町赤沼２５２０番地　株式会社日立製作所　基礎研究所内
最終頁に続く

(54)【発明の名称】複数文書データベースを対象とした文書検索方法および文書検索サービス

(57)【特許請求の範囲】
【請求項１】
キーワードもしくは文章を入力させて第１のサーバへ送信するための入力部と、前記第１のサーバから受信した検索結果を表示する検索結果表示部と、前記検索結果表示部に表示された文書群を選択させて選択された文書群の文書識別子を前記第１のサーバへ送信するための文書群選択部と、前記選択された文書群の概略を前記第１のサーバから受信して表示する概略表示部と、前記概略を第２のサーバへ送信させる概略送信機構とを有するインターフェース部と、
第１の文書群を格納する第１の文書データベースと、前記インターフェース部から受信する前記キーワードもしくは文章と関連度の高い文書群を前記第１の文書データベースから検索して、前記検索結果を前記インターフェース部へ送信させるための第１の検索機構と、前記文書群選択部によって選択された文書群の識別子を受信し、選択された文書群の概略を、前記文書群中の単語と前記単語の重みを用いて作成して、前記概略を前記インターフェース部へ送信させるための概略作成機構とを有する第１のサーバと、
第２の複数文書群を格納する第２の文書データベースと、前記インターフェース部から受信する前記概略との関連度の高い文書群を前記第２の文書データベースから検索する第２の検索機構を具備し、前記第２の検索機構による検索の結果である第２の文書識別子を、関連度重みを付加して前記検索結果表示部に表示させるよう前記インターフェース部へ送信する第２のサーバとを有することを特徴とする文書検索装置。
【請求項２】

前記概略は、前記選択した文書群中での出現頻度、及び前記第１の文書データベース中での出現頻度を入力とする関数の計算で得る重みが所定の閾値以上である単語の集合であることを特徴とする請求項１に記載の文書検索装置。

【請求項３】

前記関連度重みは、前記概略及び前記第２の文書データベース中の文書群の両方に含まれる単語につき、前記概略中の重み、前記第２の文書データベースの文書群中の重みから総合的な重み、及び前記単語の重みの集計を計算することにより得るものであることを特徴とする請求項１に記載の文書検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、複数の文書データベースを切りかえて検索したり、複数の文書データベースを関連付けたりするための文書検索方法に関する。

【０００２】

【従来の技術】

様々な文書情報が電子化されるにつれ、複数の異質な文書データベースを同時に検索する必要性が増してきている。例えば、興味を引いた新聞記事からそれに関連する百科辞典の項目を閲覧するといった要求は多い。

【０００３】

従来の検索技術においても複数の文書データベースを切りか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内の文書群に対して、別の文書データベース内の関連する文書群を検索するといった、文書データベース間に渡る文書群間の関連性を調べることはできなかった。

【０００４】

これに対し、同一文書データベースに限れば、文書群を検索入力として指定して、同じ文書データベース内の関連する文書群を検索することは可能である。この場合、検索に先立って文書間の関連度を計算しておくことにより十分な検索速度を得る場合が多い。異なる文書データベース間においても、このような前計算を行えば、複数文書データベースを同時に検索することも可能であるが、文書データベースの数が増すにつれ、前計算の必要数も組み合わせ的に増大するため、この方法も現実的には不可能である。

【０００５】

また、利用者側で一旦検索元の文書群を解析し、検索入力を構成して、他の文書データベース内を検索することも可能であるが、この場合、利用者側が検索元の文書群に関する全情報を受けとらなければならず、文書データベースが通信ネットワーク上に存在する場合、通信量が膨大になってしまう。

【０００６】

【発明が解決しようとする課題】

前記従来技術の問題を解消し、利用者が、任意の文書データベース中の任意の文書群を指定し、その文書群に関連する文書群を、更に任意の文書データベース内から効率よく検索できるようにすることである。

【０００７】

【課題を解決するための手段】

文書群のように検索入力が大きい場合、検索入力の全情報を検索に使うのではなく、検索入力内の特徴的な単語のみを概要として検索に使うことで、検索速度が速く、かつ、通信ネットワークへの負荷も小さい検索方法を実現する。

【０００８】

各文書データベースに関しては、指定された文書群に対して、その中で特徴的な単語を選択することで概要を作成する機構と、送られてくる任意の概要に対して検索を行う機構を有するサーバーとして通信ネットワークに配置する。

【０００９】

検索を行う利用者は、クライアントを介して、まず、検索元の文書データベースが格納されているサーバーに対して文書群を指定して、その概要を受けとる。次に受けとった概要を検索先の文書データベースが格納されているサーバーに送り、検索結果を受けとる。

【００１０】

クライアントの検索インターフェースとしては、まず、文書群の表示エリアを有し、このエリアにおいて必要な文書群を指定することができるようにする。また検索先のデータベースも選択できるようにする。これによりクライアントでは、文書群表示エリアに表示されている文書群の中から、利用者が興味ある文書群を選択して、必要なら検索先の文書データベースを切りかえて検索を行うことができる。

【００１１】

【発明の実施の形態】

図１は、クライアント１１がサーバ１３の文書データベース１３１内の任意の文書群を指定して、指定した文書群と関連度（類似度）の高い文書群を別のサーバ１４の文書データベース１４１から得るための方法を実現する全体構成の一例を示したものである。ここで、検索元、検索先の文書データベース１３１、１４１は通信ネットワーク１２を介してアクセスできる異なった場所にあるサーバ上にそれぞれ配置されている。

【００１２】

まず、クライアント１１は、利用者の入力に応じて検索元となる文書データベース１３１内の文書群を指定し、サーバ１３が理解できる文書識別子の集合として通信ネットワーク１２を介してサーバ１３に送出する。文書群の指定は、後述する検索結果表示部（文書群指定部）Ｐ１にて行う。

【００１３】

サーバ１３は、検索機構１３３により、クライアントから送られてきた文書識別子の集合と関連度の高い文書群を文書データベース１３１から検索する。この際、概略作成機構１３２により、検索された文書群に対して文書群の概略を作成し、通信ネットワーク１２を介してクライアント１１に返答する。ここでの概略とは、文書群をよくあらわす単語の集合のことある。概略作成機構の実施形態は特開平９－６２６９３「確率モデルによる文書分類方法」などの既存の方法が利用できる。

【００１４】

一例を示すと、まず、概略を作成しようとする文書群中の全文書を単語に分割して頻度集計する。一般に、ある文書群で良く現れる単語ほどその文書群を代表する度合も高いため、文書群中で出現頻度が高い単語ほど概略に含まれやすいことになる。ただし「する」などのように、どの文書にもよく現れるような一般語は概略として適当ではない。よって通常は、文書群が属する文書データベース中での出現頻度も考慮して概略としての単語選択を行う。つまり、指定された文書群中での出現頻度が高く、かつ、文書データベース全体での総出現頻度が低い単語ほど、その文書群中でしか現れないという意味で特徴的な単語であり、その文書群を特徴付ける概略として適切である。具体的には、文書群中のそれぞれの単語について、文書群中での出現頻度、文書データベース中での出現頻度を入力とする適当な関数により単語の重みを計算し、ある閾値以上の重みを持つ単語を概略として採用する。サーバー１３は以上の方法で作成した重み付き単語の集合を通信ネットワーク１２を介してクライアントに返す。この単語を図２では、「特徴語」として表示するものとしている。

【００１５】

次に、クライアント１１はサーバ１３から返答された概略（検索元の文書群の概略）を評価しあるいは加工して、通信ネットワーク１２を介して検索先のサーバ１４に送出する。

【００１６】

サーバ１４は、検索機構１４３により、クライアントから送られてきた文書群の概略と関連度の高い文書群を検索先の文書データベース１４１から検索し、検索結果の文書識別子を関連度の重み付きでクライアント１１に返す。ここでの検索機構は、公知のキーワード検索法により実現できる。つまり、入力である文書群概略は重み付き単語の集合であるた

め、各単語を重み付きの入力キーワードとみなしＯＲ検索すればよい。その際、検索結果の文書の重み（関連度）は以下のように計算できる。概略および検索先の文書両方に含まれる各単語について、概略中での重みと、検索先の文書における重み（例えば頻度）から総合的な重みを計算し（例えば両重みの積）、さらにそのような単語全てに関する重みを集計（例えば総和）することで関連度を得る。

【００１７】

以上の方法で、クライアント１１は文書データベース１３１内の任意の文書群に関連する文書データベース１４１内の文書群を得ることができる。ここでの特徴は、検索元の文書群に関する処理（概略作成）をサーバ側に任せることにより、通信ネットワーク中の通信量が少なくて済む点である。クライアントが検索元の文書全文情報を一旦受けとって処理する場合に比べると差は歴然である。クライアントの検索支援機構１１２では、基本的には検索元のサーバから送られてきた文書群概略を検索先サーバに送るだけでよく、検索に関わる処理のほとんどを両サーバにまかせることができる。一方サーバ側は、担当する文書データベースに関して、概略作成機構、検索機構を持つのみでよく、モジュール化されており、他の文書データベースの情報に関しては全く考慮しなくてもよい。

【００１８】

以上、文書データベース１３１を検索元として文書データベース１４１を検索する手段を説明したが、全く同様の方法で、逆に文書データベース１４１を検索元として文書データベース１３１を検索することも可能である。この場合、クライアントは、文書データベース１４１中の文書群の概略をサーバ１４の概略作成機構１４２から得て、検索先のサーバ１３に送信し、サーバ１３の検索機構１３３により文書データベース１３１中の関連する文書群を得る。以上を一般化すると、新たな文書データベースに関しては、その概略作成機構、検索機構を持つサーバを用意し、通信ネットワークに接続するだけで、その文書データベースは、通信ネットワークに接続されている全ての文書データベースに対して検索元にも検索先にもなり得る。

【００１９】

最後にクライアントに関する実施形態を図２で説明する。１１１はクライアントに搭載されている検索支援インターフェースの例で、これは基本的には、特願平９－２４０９６３「文書検索支援方法および文書検索支援サービス」で本願の発明者らによって提案されたものと同じである。Ｅ１は検索要求の入力ウィンドウであり、利用者は、ここに検索要求をキーワードの羅列、または文章形式で入力できる。Ｍ１は文書データベース選択ウィンドウであり、利用者が、右端の指示部をマウスでプルダウンすることにより文書データベースの一覧があらわれ、所望の文書データベースを選択できる。Ｂ１は検索の開始を指示するボタンである。よって、利用者はウィンドウＥ１に任意の検索要求を入力し、ウィンドウＭ１で検索対象の文書データベースを選択し、ボタンＢ１を押すことで、ウィンドウＭ１で選択した文書データベースに対してウィンドウＥ１に入力したキーワードによる通常のキーワード検索を実行させることができる。この検索の実行は図1に示す検索支援機構１１２の支援のもとに実行されるが、これの詳細は、先の出願に詳しいので、ここでは説明を省略する。もちろん、一般に行われるキーワード検索によっても良いことは言うまでもない。

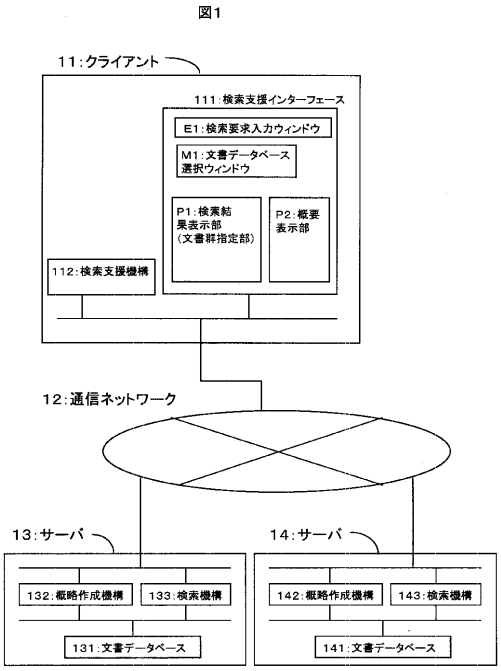
【００２０】

Ｐ１は検索結果表示部であり、上段に選択の結果選択された文書の総数および後述するようにして利用者に選択された文書の数を表示する窓が、その下に、利用者の選択／非選択を入力する窓、検索要求との関連度および選択された文書のタイトルの一覧（リスト形式で表示）を示す表示部が配置されたものとなっている。この表示部はスクロール機能を持っており表示に一度に表示できない部分もスクロールによって見ることができる。選択／非選択を入力する窓はマウスによりクリック可能であり、マウスでクリックする毎に選択／非選択の状態が反転する。これをクリックして選択にすると、この文書に対応した文書の概要が、概要表示部Ｐ２に重み付きの単語集合のグラフ形式で表示される。概要表示部Ｐ２にも、上段部に、特徴語の総数および利用者に選択された特徴語の数を表示する窓が

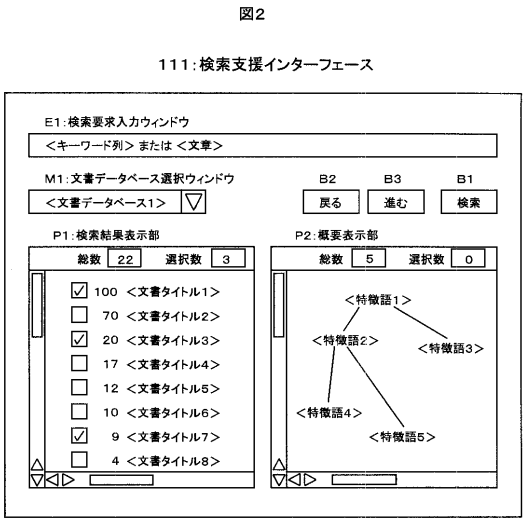
設けられる。なお、文書タイトルは、通常、関連度の順にソートされている。
【００２１】
図の検索結果表示部Ｐ１は、検索の結果選択された文書が総数で２２あり、利用者が選択された文書のタイトルから興味ある文書として三つの文書を選択を選択している状態を示す。選択された文書はクリックによってチェックマークが表示されている。概要表示部Ｐ２は、これに応じて、選択された文書の検索要求入力に対応する特徴語が５つ表示されている。
【００２２】
この実施例では説明を省略するが、概要表示部Ｐ２に表示された特徴語を選択することでこれをキーとする文書を検索結果表示部Ｐ１に逆に表示させることもできる。したがって、利用者が自分の好みにカスタマイズした概要によって、よりきめ細かい検索が可能になる。これについては、先に引用した特願平９－２４０９６３に詳しく説明されている。
【００２３】
利用者が、このように、選択された文書のタイトルを参照しながら、文書の選択／非選択を行い、興味のある文書を複数個選択することができる。
その後、利用者が、この検索結果に対応した文書群についての他の文書データでの扱い等に興味を持ったときは、ウィンドウＭ１により文書データベースを切りかえて、検索の開始を指示するボタンＢ１を押す。
【００２４】
クライアントは、これに応じて、検索元の文書データベースが格納されているサーバ(例えばサーバ１３)に選択した複数文書の識別子を送り、それら複数文書の概要を得て、検索先の文書データベースが格納されているサーバ(例えばサーバ１４)にこの概要を送り、検索先のサーバ(例えばサーバ１４)から検索結果を得る。新しい検索結果は検索結果表示部Ｐ１に表示される。つまりこの例の場合、Ｐ１は新しく検索された文書群に書きかえられる。
【００２５】
新しい検索結果と、先の検索結果とを比較するため、先の検索結果をふたたび検索結果表示部Ｐ１に表示したければ、利用者はボタンＢ２を押して、検索結果表示部Ｐ１の表示を検索前の状態に戻すことができる。同様にボタンＢ３を押して、検索結果表示部Ｐ１の表示を新しい検索結果に進めることができる。
【００２６】
このような検索結果に応じた他の文書データベースでの検索は、検索の任意の段階で実行することができるから、検索サイクルを繰り返すことにより利用者は文書データベースから文書データベースへと探索を自由に進めることができる。当然、文書データベースを切りかえずに同一文書データベース内でこのサイクルを繰り返すことも可能である。
【００２７】
【発明の効果】
利用者は、各文書データベースの配置、構成について意識すること無く、検索対象としての文書データベースを自由に指定して、探索を自由に進めることができる。また、文書データベースを保有するサーバはモジュール化できるため、新たな文書データベースに関しては、その概要作成機構、検索機構を有するサーバを通信ネットワークに接続するだけで、通信ネットワーク上の他の文書データベース全てに対して、検索先にも検索元にもなることが可能である。
【図面の簡単な説明】
【図１】複数文書データベース検索方法のシステムの全体構成の一例を示す図。
【図２】クライアントにおける検索支援インターフェースの構成の一例を示す図。
【符号の説明】
１１：クライアント、１１１：検索支援インターフェース、１１２：検索支援機構、１２：通信ネットワーク、１３：サーバ、１３１：文書データベース、１３２：概略作成機構、１３３：検索機構、１４：サーバ、１４１：文書データベース、１４２：概略作成機構

、１４３：検索機構、Ｂ１：検索ボタン、Ｂ２：戻りボタン、Ｂ３：進むボタン、Ｅ１：検索要求入力ウィンドウ、Ｍ１：文書データベース選択ウィンドウ、Ｐ１：検索結果表示部、Ｐ２：概要表示部。

【図１】



【図２】



フロントページの続き

(72)発明者　丹羽　芳樹
埼玉県比企郡鳩山町赤沼２５２０番地　株式会社日立製作所　基礎研究所内
(72)発明者　高野　明彦
埼玉県比企郡鳩山町赤沼２５２０番地　株式会社日立製作所　基礎研究所内

審査官　深津　始

(56)参考文献　特開平０４－１３８５６３（ＪＰ，Ａ）
特開平０９－２１８８８１（ＪＰ，Ａ）
特開平１０－２６９２３７（ＪＰ，Ａ）
丹羽芳樹，動的な共起解析を用いた対話的文書検索支援，情報処理学会研究報告，１９９６年９月１３日，９６巻，８号，４１－４８頁，９６－ＦＩ－４３
Michelle Q. Wang Baldonado，ＳｅｎｓｅＭａｋｅｒによる検索、ブラウズ、メタサーチ，internetworking，株式会社アスキー，１９９７年　８月　１日，３巻，８号，３１－３６頁

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30　　-G06F 17/30　419