(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3715672号
(P3715672)
(45)発行日　平成17年11月9日(2005.11.9)　(24)登録日　平成17年9月2日(2005.9.2)

(51)Int.Cl.[7]　F I
G06F 17/30　G06F 17/30　350C
G06F 17/21　G06F 17/30　170A
G06F 17/21　590E

請求項の数 7　(全 15 頁)

(21)出願番号　特願平7-41714
(22)出願日　平成7年3月1日(1995.3.1)
(65)公開番号　特開平8-235212
(43)公開日　平成8年9月13日(1996.9.13)
審査請求日　平成14年3月1日(2002.3.1)

(73)特許権者　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　100076428
弁理士　大塚　康徳
(74)代理人　100093908
弁理士　松本　研一
(72)発明者　伊藤　史朗
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内
(72)発明者　廣田　誠
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

最終頁に続く

(54)【発明の名称】テキスト検索方法及び装置

(57)【特許請求の範囲】

【請求項１】

被検索テキストから指定された検索文字列を検索するテキスト検索装置であって、

被検索テキスト中の各文字出現位置のインデックスを保持するインデックス保持手段と、

指定された検索文字列中の各文字が被検索テキスト中に出現する位置を前記インデックス保持手段から取得するインデックス取得手段と、

前記取得した被検索テキスト中に出現する各文字の各位置に基づいて、前記検索文字列中の各文字と出現位置関係が一致もしくは所定範囲内の位置にある文字の個数をカウントし、該個数に基づいて、前記被検索テキスト中に出現する前記各文字を含んで構成される各被検索文字列の類似度を獲得する類似度獲得手段と、

前記獲得手段で獲得された類似度が所定の閾値以上の場合、前記被検索文字列を検索結果として決定する決定手段とを備えることを特徴とするテキスト検索装置。

【請求項２】

前記類似度獲得手段は、前記検索文字列中で連続する文字が前記被検索文字列中で連続して一致する場合に、類似度を加算することを特徴とする請求項１記載のテキスト検索装置。

【請求項３】

検索文字列の各文字の位置関係と前記インデックス保持手段に保持された各文字の位置関係とが完全に一致する被検索文字列を抽出する完全一致文字列抽出手段をさらに備える

ことを特徴とする請求項１記載のテキスト検索装置。
【請求項４】
完全一致検索を行うかあいまい検索を行うかの検索状態を保持する検索状態保持手段を更に備え、
前記検索状態保持手段に保持された検索状態があいまい検索を指示する場合は前記類似度獲得手段と決定手段を実行し、該検索状態が完全一致検索を指示する場合は前記完全一致文字列抽出手段を実行することを特徴とする請求項１乃至３記載のテキスト検索装置。
【請求項５】
前記閾値を保持する閾値保持手段を更に備え、
前記閾値保持手段に保持された閾値に応じてあいまい検索と完全一致検索を切り替えて実行することを特徴とする請求項１乃至４記載のテキスト検索装置。
【請求項６】
前記類似度獲得手段で獲得された類似度の順に結果を保持する検索結果保持手段を更に備えることを特徴とする請求項１又は２に記載のテキスト検索装置。
【請求項７】
テキスト記憶手段に格納された被検索テキストから指定された検索文字列を検索するテキスト検索装置によるテキスト検索方法であって、
前記テキスト検索装置のインデックス取得手段が、被検索テキスト中の各文字出現位置のインデックスを保持するインデックス保持手段から、指定された検索文字列の各文字が被検索テキストに出現する位置を取得するインデックス取得工程と、
前記テキスト検索装置の類似度獲得手段が、前記取得した被検索テキスト中に出現する各文字の各位置に基づいて、前記検索文字列中の各文字と出現位置関係が一致もしくは所定範囲内の位置にある文字の個数をカウントし、該個数に基づいて、前記被検索テキスト中に出現する前記各文字を含んで構成される各被検索文字列の類似度を獲得する類似度獲得工程と、
前記テキスト検索装置の決定手段が、前記類似度獲得工程で獲得された類似度が閾値記憶手段に格納されている所定の閾値以上の場合、前記被検索文字列を検索結果として決定する決定工程とを備えることを特徴とするテキスト検索方法。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】
本発明は、インデックスを用いてテキストデータから文字列を高速に検索するテキスト検索方法及び装置に関するものである。
【０００２】
【従来の技術】
文書中の全てのテキストを対象として与えられた検索語を含む文書を検索する全文検索装置などのテキスト検索装置では、大量のテキストを高速に検索するために、被検索テキストのインデックスを予め作成して、このインデックスを用いて検索を行なうインデックス技術が利用されている。インデックス技術の一例として、特開平４－２０５５６０号公報では、文字位置インデックス技術が述べられている。
【０００３】
文字位置インデックス技術の基本的な考え方は、被検索テキスト中に出現する文字および文字列の位置を文字ごとに１ずつ増加する整数で表わすことにある。その上で、各文字および文字列ごとに、当該文字および文字列をキーとして、当該文字および文字列が現れる全ての位置を列挙する。例えば、図２に示す文書に対して、図３に示すインデックスを作成する。図３の各行は、それぞれの文字が現れる位置を列挙したものである。例えば、文字「高」は、位置０，…，０ｘ０２ａ７４ｃ５ｆ，０ｘ０２ａ７４ｃ６５，…に現れている。
【０００４】
このインデックスにおいて、ある検索文字列を被検索テキストから検索する場合には、当

該検索文字列をインデックスのキーとなっている文字および文字列に分解し、分解した文字および文字列の位置関係が、当該検索文字列における位置関係に一致する組合せを探すことで検索を行なう。このように、文字位置インデックス技術は、検索文字列と完全に一致する文字列を検索する技術である。

【０００５】

一方、テキスト検索装置では、検索文字列として語が与えられた場合、その異表記語や同義語などをも検索する必要性が高い。これを実現する技術の一つとして、与えられる検索文字列に表層的に類似する文字列を検索するあいまい検索技術がある。

【０００６】

特に、指定する文字列に表層的に類似する文字列を検索するあいまい検索を行なうテキスト検索装置に関するものである。

【０００７】

あいまい検索技術の一例として、北陸先端科学技術大学院大学情報科学研究科Research Report IS-RR-936I, 「用例検索による日英翻訳支援システムＣＴＭ２－部分文字列インデックスを用いた最適照合検索」において示されているあいまい検索技術がある。このあいまい検索技術では、検索文字列を長さｋの部分文字列に分解する。次に、各部分文字列が被検索テキスト内の特定ブロックに存在する場合に、そのブロックのスコアに１点加点をすることで、スコアが高いブロックを検索するものである。ｋは任意であるが、上記の文献では、ｋが２と３の２種類で行い、その合計値をスコアとするのが適当であると示されている。

【０００８】

また、別のあいまい検索技術の一例として、情報処理学会研究報告, 情報学基礎Vol.92 No.32 pp.27-32(1992)に掲載されている「高速全文不完全一致検索システムの実現」において示されているあいまい検索技術がある。

【０００９】

上述したいずれのあいまい検索技術も、検索文字列に表層的に近い文字列を検索する技術であって、検索文字列に完全に一致する文字列だけを検索することはできない。

【００１０】

【発明が解決しようとする課題】

上述した様に、従来例の技術によるテキスト検索装置では、インデックスは完全一致検索用かあいまい検索用のいずれかに特化されている。従って、一つのインデックスを用いた場合は、検索文字列と完全に一致する文字列を検索する完全一致検索か、検索文字列と表層的に近い文字列を検索するあいまい検索かいずれか一方の検索しか行うことができなかった。完全一致検索とあいまい検索とを同時に行なうためには、二つのインデックスを作成しなければならず、インデックスサイズの増大やインデックス作成時間の増大を招くという欠点があった。

【００１１】

本発明は上記の問題点に鑑みてなされたものであり、文字位置を表す文字位置インデックスを用いてあいまい検索を行うことが可能なテキスト検索方法及び装置を提供することを目的とする。

【００１２】

また、本発明の他の目的は、あいまい検索と完全一致検索とを同一のインデックスを用いて行うことを可能とし、インデックスの大きさ及びその作成処理時間を低減するテキスト検索方法及び装置を提供することにある。

【００１３】

【課題を解決するための手段】

上記の目的を達成するための本発明のテキスト検索装置は以下の構成を備えている。即ち、

被検索テキストから指定された検索文字列を検索するテキスト検索装置であって、

被検索テキスト中の各文字出現位置のインデックスを保持するインデックス保持手段と

、
指定された検索文字列中の各文字が被検索テキスト中に出現する位置を前記インデックス保持手段から取得するインデックス取得手段と、
前記取得した被検索テキスト中に出現する各文字の各位置に基づいて、前記検索文字列中の各文字と出現位置関係が一致もしくは所定範囲内の位置にある文字の個数をカウントし、該個数に基づいて、前記被検索テキスト中に出現する前記各文字を含んで構成される各被検索文字列の類似度を獲得する類似度獲得手段と、
前記獲得手段で獲得された類似度が所定の閾値以上の場合、前記被検索文字列を検索結果として決定する決定手段とを備える。
【００１４】
また、上記の目的を達成する本発明の他の態様によるテキスト検索方法は、
テキスト記憶手段に格納された被検索テキストから指定された検索文字列を検索するテキスト検索装置によるテキスト検索方法であって、
前記テキスト検索装置のインデックス取得手段が、被検索テキスト中の各文字出現位置のインデックスを保持するインデックス保持手段から、指定された検索文字列の各文字が被検索テキストに出現する位置を取得するインデックス取得工程と、
前記テキスト検索装置の類似度獲得手段が、前記取得した被検索テキスト中に出現する各文字の各位置に基づいて、前記検索文字列中の各文字と出現位置関係が一致もしくは所定範囲内の位置にある文字の個数をカウントし、該個数に基づいて、前記被検索テキスト中に出現する前記各文字を含んで構成される各被検索文字列の類似度を獲得する類似度獲得工程と、
前記テキスト検索装置の決定手段が、前記類似度獲得工程で獲得された類似度が閾値記憶手段に格納されている所定の閾値以上の場合、前記被検索文字列を検索結果として決定する決定工程とを備える。
【００２０】
【実施例】
以下、添付の図面を参照して本発明の好適な実施例を詳細に説明する。
【００２１】
＜実施例１＞
図１は、実施例１のテキスト検索装置の基本構成を示すブロック図である。同図において、１０１は文字位置インデックス保持部であり、被検索テキスト中の文字および文字列をキーとして、被検索テキスト中での該当文字の位置を保持したインデックスを保持する。１０２は検索語保持部であり、検索を行なう文字列を保持する。１０３は類似度閾値保持部であり、被検索文字列を検索結果とするか否かを判断するための類似度閾値を保持する。検索文字列と被検索文字列との表層的な類似度がこの類似度閾値より高い場合には、当該被検索文字列を検索結果とする。
【００２２】
１０４はあいまい検索部であり、文字位置インデックス保持部１０１に保持されている文字位置インデックスを用いて、被検索テキスト中の任意の文字列について、検索文字列保持部１０２に保持されている検索文字列との表層的な類似度を計算する。更に、あいまい検索部１０４は、算出された類似度が類似度閾値保持部１０３に保持されている類似度閾値を越える文字列を検索する。１０５は検索結果保持部であり、あいまい検索部１０４による検索結果を保持する。
【００２３】
あいまい検索部１０４における類似度の計算方法には、様々なものが考えられるが、ここではその一例を示す。検索文字列の長さをＬとした場合、類似度の閾値ａを０＜ａ＜Ｌの範囲で設定する。仮想位置ｐを設定した場合に、検索文字列を分解した各文字の位置ｐｉ（１≦ｉ≦Ｌ）で、
　－（Ｌ－ａ）≦（ｐｉ－ｉ）－ｐ≦（Ｌ－ａ）
を満足する個数を類似度とする。仮想位置ｐに対する類似度が、ａ以上であれば、その仮

想位置近辺にある文字列を検索結果とする。

【００２４】

次に、文字位置インデックス保持部１０１に保持されている文字位置インデックスについて説明する。図２は文書の一例を表す図である。また、図３は図２に示された文書に対して生成された文字位置インデックスを表す図である。図３に示された文字位置インデックスの各行は、それぞれの文字が現れる位置を列挙したものである。例えば、文字「高」は、位置０，…，０ｘ０２ａ７４ｃ５ｆ，０ｘ０２ａ７４ｃ６５，…に現れていることがわかる。

【００２５】

図４は実施例１のテキスト検索装置の制御構成を表すブロック図である。同図において、１はＣＰＵであり、ＲＯＭ２に格納された制御プログラムを実行することにより各種の制御を実現する。２はＲＯＭであり、ＣＰＵ１によって実行される各種の制御プログラムを格納する。ＣＰＵ１とＲＯＭ２は、特に、後述するフローチャートで示される制御を実行することにより、図１で示した各部の機能を実現する。３はＲＡＭであり、各種データを一時的に保持する。ＲＡＭ３は、文字位置インデックス保持部１０１、検索語保持部１０２、類似度閾値保持部１０３、検索結果保持部１０５のための格納領域を提供する。

【００２６】

４は入力部であり、検索語の入力や、類似度設定のためのデータ入力等を行う。５はディスプレイであり、文書の表示や、検索結果の表示を行う。６は外部記憶装置であり、検索対象となる文書データ等を格納する。７はバスであり、図４に示された上記の各構成を接続する。

【００２７】

図５は、実施例１のテキスト検索装置による検索処理の手順を表すフローチャートである。本フローチャートで示される制御をＣＰＵ１に実現させるための制御プログラムはＲＯＭ２に保持されている。以下、図５のフローチャートを参照して、上記の類似度計算例に従った本装置の処理を説明する。

【００２８】

ステップＳ２０１では、検索文字列保持部１０２に保持されている検索文字列の長さをＬに代入する。また、ｎには初期値として１を代入する。例えば、検索文字列が「高速探索」である場合は、Ｌ＝４，ｎ＝１となる。また、類似度閾値保持部１０３に保持されている類似度の閾値をａに代入する。そしてステップＳ２０２に移る。

【００２９】

ステップＳ２０２では、検索文字列保持部１０２に保持されている検索文字列の１番目の文字（例えば「高速探索」であれば「高」）について、その全出現位置を文字位置インデックスより読み込む。読み込まれた文字位置全てを第１配列に格納する。図６は実施例１による第１配列のデータ格納状態を示す図である。ここで、第１配列の列１００１には各文字位置を格納し、列１００２には全て１が格納される。図６においては、図３に示した文字位置インデックスを用いて検索文字列「高速探索」を検索しているときの、ステップＳ２０２の終了時点における第１配列のデータ格納状態が示されている。

【００３０】

次に、ステップＳ２０３では、変数Ｌとｎを比較し、ｎ＜Ｌである場合には、ステップＳ２０４に移る。一方、ｎ≧Ｌである場合には、ステップＳ２０７に移る。そして、ステップＳ２０４でｎの値を１増やす。

【００３１】

ステップＳ２０５では、検索文字列保持部１０２に保持されている検索文字列のｎ番目の文字の文字位置について、文字位置インデックスよりの読み込みを行ない、第２配列に格納する。このとき、文字位置インデックスより得られた当該文字の全ての文字位置から（ｎ－１）を減じた値を第２配列に格納する。第２配列は、例えば図６に示した列１００１のみで構成される配列である。

【００３２】

次に、ステップＳ２０６では、第２配列中の各値について、次の処理を行なう。即ち、

(1) 第１配列の列１００１に第２配列中の値と同じ値が存在する場合は、第１配列の当該値を有する行の列１００２の値を１増やす.

(2) 第１配列の列１００１に第２配列中の値との差がＬ－ａ以内である値がある場合は、第１配列の当該値を有する行の列１００２の値を１増やすと共に、第１配列に新たな行を追加し、この追加された行の列１００１の値を第２配列中の値に、列１００２の値を１に夫々セットする.

(3) 上記の何れでもない場合は、第１配列に新たに行を加え、当該行の列１００１の値を第２配列中の値に、列１００２の値を１にする。

【００３３】

以上のステップＳ２０６の処理を終了するとステップＳ２０３に戻り、ｎ＜Ｌを満足しなくなるまで（即ち、検索文字列の全ての文字について上述の処理が実行されるまで）上述の処理を繰り返す。本例ではｎ＝４となった時点で処理がステップＳ２０３からステップＳ２０７へ分岐する。図７は、類似度閾値を３とした場合の本例の検索処理において、ｎ＝４のときの第１配列の格納状態を示す図である。

【００３４】

ステップＳ２０７では、第１配列の列１００２の値がａ以上である行の列１００１の値を、被検索テキスト中の検索文字列の先頭文字の位置として、検索結果保持部１０５に保持する。図１１の例では、０２ａ７４ｃ６５が検索文字列の先頭文字の位置として検索結果保持部１０５に保持される。そして本処理を終了する。

【００３５】

次に、文字位置インデックス保持部１０１に保持された文字位置インデックスを用いて完全一致検索を行う場合について説明する。

【００３６】

図８は、本実施例のテキスト検索装置における完全一致検索を行うための基本構成を示すブロック図である。同図において、上記図１と同様の機能を有する構成については図１と同一の参照番号を付し、ここでは説明を省略する。４０３は検索部であり、文字位置インデックス保持部に保持されている文字位置インデックスを用いて、検索文字列保持部に保持されている検索文字列に一致する被検索テキスト中の文字列を検索する。

【００３７】

図９は、本実施例における完全一致検索の制御手順を表すフローチャートである。ここで、ステップＳ５０１～ステップＳ５０５の各処理は図５のステップＳ２０１～ステップＳ２０５の各処理と同様である。ただし、本処理で用いる第１配列は、図１０に示すように、図６の第１列１００１に相当する部分のみを有する。図１０は、図３に示した文字位置インデックスを用いて検索文字列「高速検索」を検索している時の第１配列の状態を示す図である。

【００３８】

ステップＳ５０６では、第１配列と第２配列から、第１配列と第２配列の両方に存在している値を全て取り出し、これらの値を新たに第１配列の値とする。そしてステップＳ５０３に戻る。図１１は、図３に示した文字インデックスにおいて、「高速検索」という文字列を検索した場合におけるｎ＝３のときの第１配列の状態を示す図である。

【００３９】

そして、ステップＳ５０７では、第１配列の値を被検索テキスト中の検索文字列の先頭文字の位置として、検索結果保持部１０４に保持する。そして全ての処理を終了する。この処理の結果、先の例では検索文字列「高速」を検索すると、位置０ｘ０２ａ７４ｃ６５にある文字が検索された文字列の先頭文字として検出される。

【００４０】

以上説明したように、本実施例１によれば、文字位置インデックスを用いてあいまい検索を行うことが可能となる。また、その文字位置インデックスは、そのまま完全一致検索に用いることも可能となる。

【００４１】
＜実施例２＞
上記実施例１では、検索文字列を分解した各文字がある範囲内に存在する個数を類似度とする場合について説明したが、これに限定されるものではなく、文字位置インデックスが持つ各文字の位置を用いて計算できる任意の類似度計算方法を用いてよい。ここでは、他の類似度計算の例として、文字が連続して一致する場合の類似度を高くする計算方法について説明する。尚、実施例２のテキスト検索装置の構成は実施例１（図１、図２）と同様であるので、ここではその説明を省略する。
【００４２】
実施例２において、あいまい検索部１０４は次のように処理を行う。実施例２の方法では、任意の正の係数ｋを導入する。検索文字列の長さをＬとしたとき、類似度の閾値ａを０＜ａ≦（ｋ＋１）・（Ｌ－１）の範囲で設定する。仮想位置ｐを設定した時に、検索文字列を分解した各文字の位置ｐｉ（１≦ｉ≦Ｌ）で、
－（Ｌ－ａ／（ｋ＋１））≦（ｐｉ－ｉ）－ｐ≦（Ｌ－ａ／（ｋ＋１））
を満足する個数と、上記の条件を満足するｐｊ，ｐｊ+1（１≦ｊ≦Ｌ－１）があって、
ｐｊ＝ｐｊ+1－１
を満足する組の個数にｋを乗じたものの和を類似度とする。そして、仮想位置ｐに対する類似度が、ａ以上であれば、その仮想位置近辺にある文字列を検索結果とする。
【００４３】
図１２は実施例２のテキスト検索装置の動作手順を表すフローチャートである。以下、図１２のフローチャートを参照して、上記の類似度計算例に従った実施例２の処理を説明する。
【００４４】
ステップＳ３０１では、検索文字列保持部１０２に保持されている検索文字列の長さをＬに代入する。また、ｎには初期値として１を代入する。例えば、検索文字列が「高速探索」である場合は、Ｌ＝４，ｎ＝１となる。また、ａには、類似度閾値保持部１０３に保持されている類似度の閾値が代入される。
【００４５】
次に、ステップＳ３０２では、検索文字列保持部１０２に保持されている検索文字列の１番目の文字について、その文字位置を文字位置インデックスより読み込む。そして、当該文字の文字位置全てを第１配列に読み込む。図１３は、実施例２による第１配列のデータ格納状態を表す図である。第１配列の列１２０１には文字位置インデックスより読み出された文字位置が格納される。列１２０２及び列１２０３は全て１にする。尚、列１２０２、１２０３の夫々の役割は、以降の説明で明らかとなる。そしてステップＳ３０３に移る。尚、図１３においては、図３に示した文字位置インデックスを用いて検索文字列「高速探索」を検索しているときの第１配列の状態を示している。
【００４６】
ステップＳ３０３では、変数Ｌとｎを比較し、ｎ＜Ｌである場合は、ステップＳ３０４に移る。一方、ｎ≧Ｌである場合は、ステップＳ３０７に移る。ｎ＜Ｌの場合は、検索文字列に未処理の文字が存在することを意味し、この場合処理はステップＳ３０４へ進み、ｎの値を１増やす。
【００４７】
ステップＳ３０５では、検索文字列保持部１０２に保持されている検索文字列のｎ番目の文字の文字位置を文字位置インデックスより読み込む。そして、文字位置インデックスより読み込まれた当該文字の全ての文字位置から（ｎ－１）を減じた値を第２配列に格納する。
【００４８】
ステップＳ３０６では、あいまい検索部１０４が、第２配列中の各値について、次の処理を行なう。即ち、
(1)　第１配列の列１２０１に第２配列中の値と同じ値が存在する場合は、第１配列の

当該値を有する行の列１２０２の値を１増やす.
(2)　(1)の場合において、当該行の列１２０３の値が（ｎ－１）である場合は、当該行の列１２０２の値を更にｋ増やす。その後、いずれの場合においても（即ち、当該行の列１２０３の値が（ｎ－１）であってもなくも）、当該行の列１２０３の値をｎに更新する.
(3)　第１配列の列１２０１に第２配列中の値との差がＬ－ａ／（ｋ＋１）以内である値がある場合は、第１配列の当該値を有する行の列１２０２の値を１増やすと共に列１２０３の値をｎにセットし、第１配列に新たな行を追加し、この追加された行の列１２０１の値を第２配列中の値に、列１２０２の値を１に、列１２０３の値をｎそれぞれセットする.
(4)　上記の何れでもない場合は、第１配列に新たに行を加え、当該行の列１２０１の値を第２配列中の値に、列１２０２の値を１に、列１２０３の値をｎにする。

【００４９】
以上の処理を行って処理はステップＳ３０３に戻る。図１４は、類似度閾値を４、係数ｋを１とした場合の本例の検索処理において、ｎ＝４のときの第１配列の格納状態を示す図である。

【００５０】
ステップＳ３０７では、第１配列の列１２０２の値がａ以上である行の列１２０１の値を、被検索テキスト中の検索文字列の先頭文字の位置として、検索結果保持部１０５に保持する。そして全ての処理を終了する。

【００５１】
ステップＳ３０７では、第１配列の列１２０２の値がａ以上である行の列１２０１の値を、被検索テキスト中の検索文字列の先頭文字の位置として、検索結果保持部１０５に保持する。図１４の例では、０２ａ７４ｃ６５が検索文字列の先頭文字の位置として検索結果保持部１０５に保持される。そして本処理を終了する。

【００５２】
以上説明したように、本実施例２によれば、検索文字列の一部と同じように文字が連続した場合に、高い類似度が与えられるので、連続性を考慮したあいまい検索が可能となる。尚、実施例２で用いている文字位置インデックステーブルは実施例１のものと同様であるので、実施例１と同様に完全一致検索と共用することが可能である。

【００５３】
＜実施例３＞
上記実施例１及び実施例２では、文字位置インデックスを用いてあいまい検索を行なう場合について説明し、更に、あいまい検索で用いた文字位置インデックステーブルをそのまま完全一致検索に用いることが可能であることを説明した。実施例３では、あいまい検索と、完全一致検索との動作の切り替え手法について説明する。

【００５４】
図１５は、実施例３によるテキスト検索装置の基本構成を示すブロック図である。同図において、図１と同様の機能を有する構成については、図１と同一の参照番号を付し、ここでは詳細な説明を省略する。

【００５５】
１４０４は検索状態保持部であり、完全一致検索を行なうかあいまい検索を行なうかを指定する情報をを保持する検索状態保持部である。１４０５はあいまい検索部であり、検索状態保持部１４０４に保持されている情報があいまい検索を指定している場合に、上記実施例１或は２で示したようなあいまい検索を実行する（文字位置インデックス保持部１０１に保持されている文字位置インデックスを用いて、検索文字列保持部１０２に保持されている検索文字列と被検索テキスト中の任意の文字列との表層的な類似度を計算して、類似度閾値保持部１０３に保持されている類似度閾値を越える類似度を持つ文字列を検索する）。

【００５６】

１４０６は完全一致検索部であり、検索状態保持部１４０４に保持されている情報が完全一致検索を指定している場合に、実施例１で説明し多様な完全一致検索を実行する（文字位置インデックス保持部１０１に保持されている文字位置インデックスを用いて、検索文字列保持部１０２に保持されている検索文字列に完全に一致する文字列を被検索テキストから検索する）。１４０７は検索結果保持部であり、あいまい検索部１４０５による検索結果と完全一致検索部１４０６による検索結果とを保持する。

【００５７】

［実施例４］

上記実施例３では、あいまい検索と完全一致検索の切り替えを検索状態保持部の情報に応じて切り替える場合について説明したが、これに限定されるものでなく、類似度閾値保持部に保持されている類似度の閾値の値に応じて切り替えてもよい。

【００５８】

図１６は実施例４におけるテキスト検索装置の機能構成を表すブロック図である。同図において、図１と同様の機能を有する構成には図１と同一の参照番号を付し、ここでは詳細な説明を省略する。

【００５９】

図１６において、１５０４はあいまい検索部であり、類似度閾値保持部１０３に保持されている類似度の閾値が一定の範囲内にある場合に実施例１或は実施例２で示したようなあいまい検索を実行する（文字位置インデックス保持部１０１に保持されている文字位置インデックスを用いて、検索文字列保持部１０２に保持されている検索目列と被検索テキスト中の任意の文字列との表層的な類似度を計算して、類似度閾値保持部１５０３に保持されている類似度閾値を越える類似度を持つ文字列を検索する）。１５０５は完全一致検索部であり、類似度閾値保持部１０３に保持されている類似度の閾値が上記一定の範囲の外にある場合に、実施例１で説明したような完全一致検索を実行する（文字位置インデックス保持部１０１に保持されている文字位置インデックスを用いて、検索文字列保持部１０２に保持されている検索文字列に完全に一致する文字列を被検索テキストから検索する）。１５０６は検索結果保持部であり、あいまい検索部１５０４による検索結果と完全一致検索部１５０５による検索結果とを保持する。

【００６０】

尚、上記各実施例では、検索結果を保持する順序が任意である場合について説明したが、これに限定されるものでなく、あいまい検索部で計算された類似度の順に保持するなどしてもよい。

【００６１】

また、上記各実施例では、類似度の閾値として類似度閾値保持部１０３に保持されている値を用いる場合について説明したが、これに限定されるものではなく、予め規定された閾値を用いてもよい。

【００６２】

また、上記各実施例では、インデックスのキーとして１文字を用いる場合について説明したが、これに限定されるものではなく、２文字以上の文字列をキーとするインデックスを合わせ持った文字位置インデックスを用いてもよい。

【００６３】

また、上記各実施例では、検索文字列中における文字の位置関係に基づいて、文字位置インデックスからの文字位置の読み込み時に文字位置の補正を行う（例えば図５のステップＳ２０５）が、これに限定されるものでなく、マージの計算時に文字位置の補正を考慮するなど、任意の箇所で補正を考慮してもよい。

【００６４】

また、上記各実施例では、文字位置インデックスの読み込みを各文字について一括して行なう場合について説明したが、これに限定されるものでなく、位置計算をしながら順次読み込むようにしてもよい。

【００６５】

また、上記各実施例では、文字位置インデックスの読み込みを１文字ごとに行なう場合について説明したが、これに限定されるものでなく、位置計算の前に、全ての文字に対してインデックスを読み込んでもよい。

【００６６】

また、上記各実施例では、文字位置インデックスを用いた検索において、検索語の先頭文字から順次処理を行なう場合について説明したが、これに限定されるものでなく、検索語中の任意の位置の文字から検索を行なってもよい。

【００６７】

また、上記各実施例では、検索結果として文字位置を保持する場合について説明したが、これに限定されるものでなく、検索箇所を含むブロックを保持するなど、検索された文字列に関する任意の情報を保持してよい。

【００６８】

以上説明したように、上記の各実施例によれば、文字位置インデックスを用いてあいまい検索を行なうことができ、かつ、あいまい検索と完全一致検索を同一のインデックスを用いて行なうことができるので、インデックスの大きさを削減することができると共に、インデックスの作成時間を削減することができる。

【００６９】

なお、上述の各実施例では、文字をキーとしたインデックスを用いて説明したが、公知の文字列をキーとしたインデックスを利用することも可能である。但し、上記実施例のあいまい検索では、文字をキーとしたインデックスのみが用いられる。従って、あいまい検索においては、２文字或は３文字の連続を調べるような場合に文字列をキーとしたインデックスが利用される。例えば、検索文字列中のある２文字は必ず連続しているという条件であいまい検索を行うような場合が考えられる。

【００７０】

尚、本発明は、複数の機器から構成されるシステムに適用しても１つの機器からなる装置に適用しても良い。また、本発明はシステム或いは装置に本発明により規定される処理を実行させるプログラムを供給することによって達成される場合にも適用できることはいうまでもない。

【００７１】

【発明の効果】

以上説明したように、本発明によれば、文字位置を表す文字位置インデックスを用いてあいまい検索を行うことが可能となる。このため、あいまい検索と完全一致検索とを同一のインデックスを用いて行うことが可能となり、インデックスの大きさ及びその作成処理時間が低減される。

【００７２】

【図面の簡単な説明】

【図１】実施例１のテキスト検索装置の基本構成を示すブロック図である。

【図２】文書の一例を表す図である。

【図３】図２に示された文書に対して生成された文字位置インデックスを表す図である。

【図４】実施例１のテキスト検索装置の制御構成を表すブロック図である。

【図５】実施例１のテキスト検索装置による検索処理の手順を表すフローチャートである。

【図６】実施例１による第１配列のデータ格納状態を示す図である。

【図７】実施例１の第１配列の格納状態を示す図である。

【図８】本実施例のテキスト検索装置における完全一致検索を行うための基本構成を示すブロック図である。

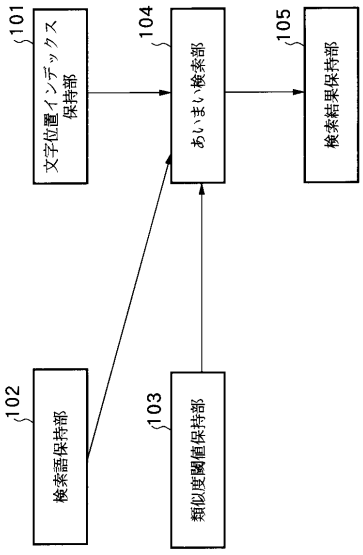
【図９】本実施例における完全一致検索の制御手順を表すフローチャートである。

【図１０】完全一致検索時における第１配列のデータ格納状態を示す図である。

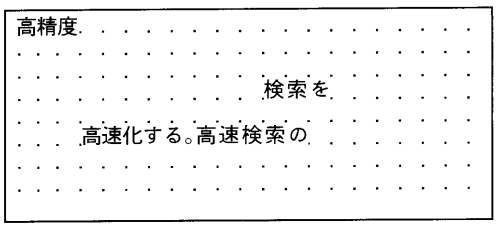
【図１１】図３に示した文字位置インデックスを用いた完全一致検索において、「高速検索」という文字列を検索した場合のｎ＝３のときの第１配列の状態を示す図である。

【図１２】実施例２のテキスト検索装置の動作手順を表すフローチャートである。
【図１３】実施例２による第１配列のデータ格納状態を表す図である。
【図１４】実施例２による検索処理において、ｎ＝４のときの第１配列の格納状態を示す図である。
【図１５】実施例３によるテキスト検索装置の基本構成を示すブロック図である。
【図１６】実施例４におけるテキスト検索装置の機能構成を表すブロック図である。
【符号の説明】
１０１　文字位置インデックス保持部
１０２　検索語保持部
１０３　類似度閾値保持部
１０４　あいまい検索部
１０５　検索結果保持部

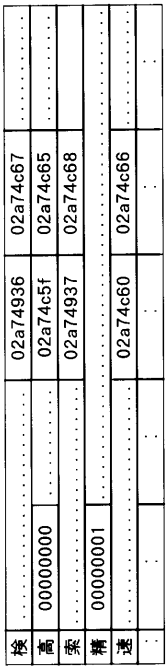
【図１】



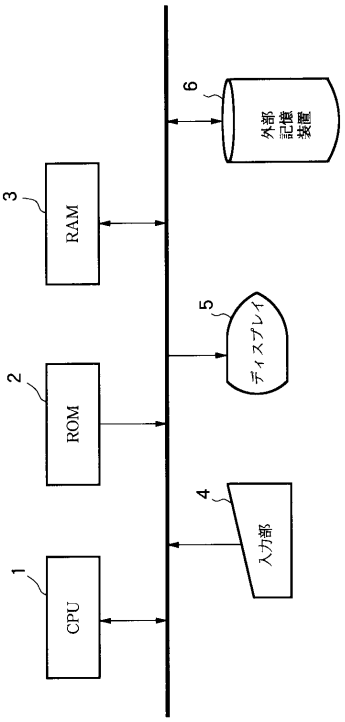
【図２】



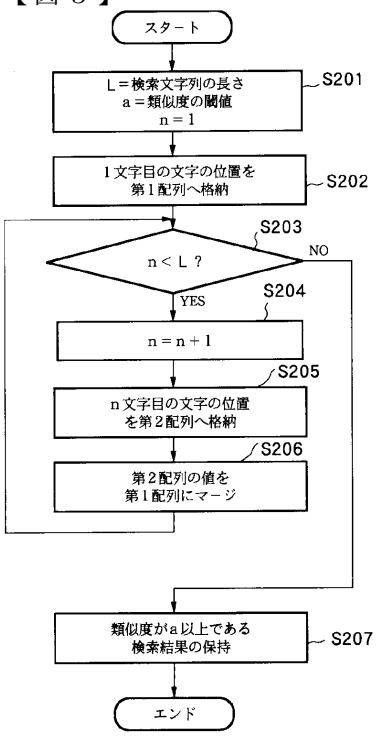
【図３】



【図４】



【図５】



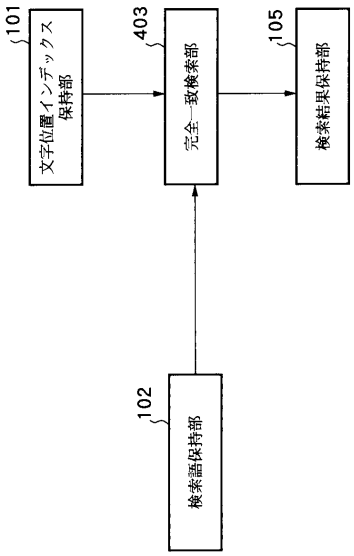
【図６】

1001	1002
00000000	1
:	:
02a74c5f	1
02a74c65	1
:	:

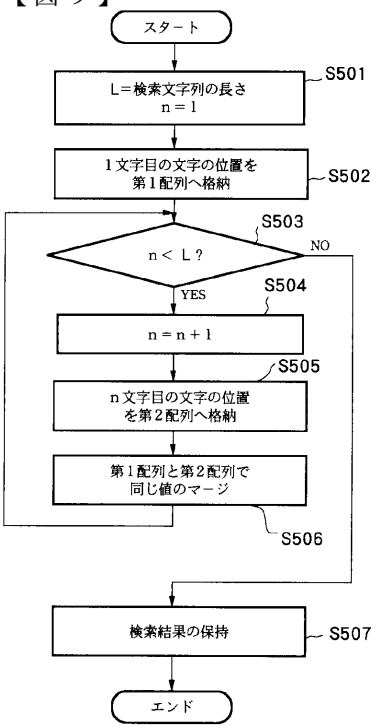
【図７】

00000000	1
:	:
02a74934	1
02a74c5f	2
02a74c65	3
:	:

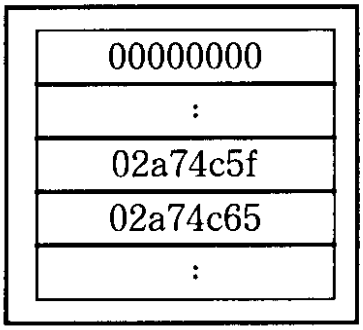
【図８】



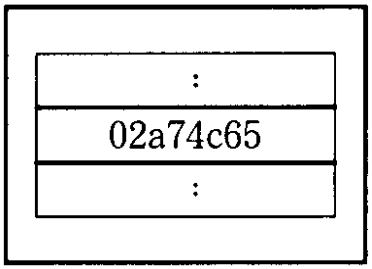
【図９】



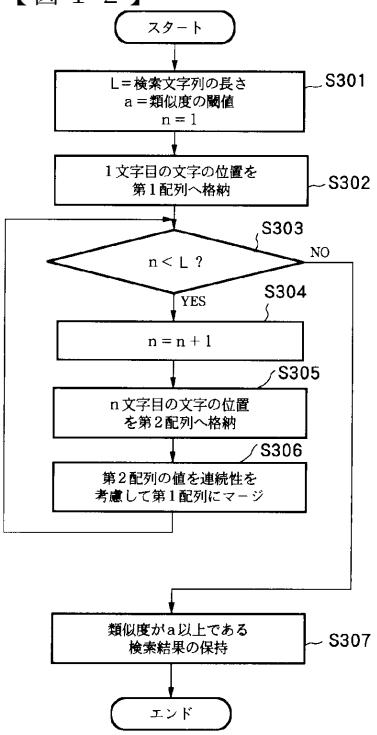
【図１０】



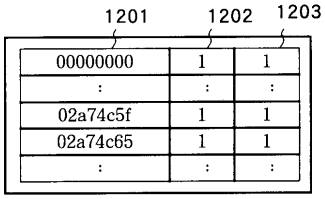
【図１１】



【図１２】



【図１３】

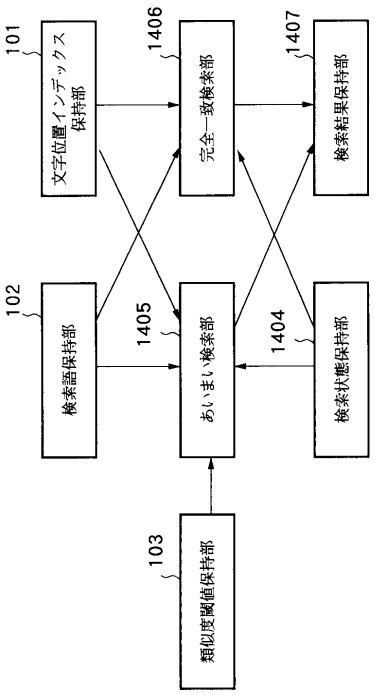


00000000	1	1
:	:	:
02a74c5f	1	1
02a74c65	1	1
:	:	:

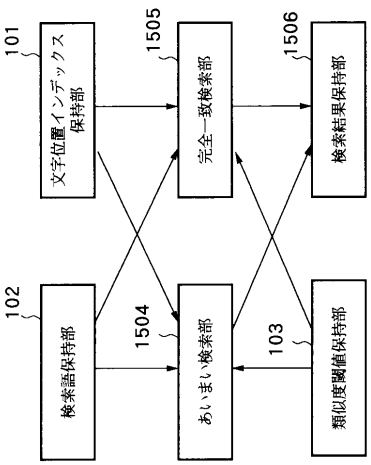
【図１４】

00000000	1	1
:	:	:
02a74934	1	4
02a74c5f	3	2
02a74c65	4	4
:	:	:

【図１５】



【図１６】



フロントページの続き

(72)発明者　柴田　昇吾
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内
(72)発明者　上田　隆也
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内
(72)発明者　池田　裕治
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内
(72)発明者　藤田　稔
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

審査官　深津　始

(58)調査した分野(Int.Cl.[7], ＤＢ名)
G06F 17/30　-G06F 17/30　419
G06F 17/21　590