(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5240475号
(P5240475)
(45)発行日　平成25年7月17日(2013.7.17)　(24)登録日　平成25年4月12日(2013.4.12)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 170A
G06F 17/30 350C
G06F 17/30 415

請求項の数 30　(全 21 頁)

(21)出願番号　特願2009-509952 (P2009-509952)
(86)(22)出願日　平成19年4月24日 (2007.4.24)
(65)公表番号　特表2009-535747 (P2009-535747A)
(43)公表日　平成21年10月1日 (2009.10.1)
(86)国際出願番号　PCT/US2007/067319
(87)国際公開番号　WO2007/130818
(87)国際公開日　平成19年11月15日 (2007.11.15)
審査請求日　平成22年4月22日 (2010.4.22)
(31)優先権主張番号　11/381,214
(32)優先日　平成18年5月2日 (2006.5.2)
(33)優先権主張国　米国 (US)

(73)特許権者　508137947
エクセジー・インコーポレイテツド
アメリカ合衆国、ミズーリ・６３１２７、セント・ルイス、サウス・ガイヤー・ロード・３６６８、スイート・３００
(74)代理人　100103920
弁理士　大崎　勝真
(74)代理人　100114188
弁理士　小野　誠
(74)代理人　100140523
弁理士　渡邊　千尋
(74)代理人　100119253
弁理士　金山　賢教
(74)代理人　100124855
弁理士　坪倉　道明

最終頁に続く

(54)【発明の名称】近似パターン合致の方法および装置

(57)【特許請求の範囲】
【請求項１】
複数のデータシンボルを含むデータ文字列が複数のパターンのうちのいずれかとの近似合致を含むか否かの決定を容易化するために、データ文字列内の複数のデータサブ文字列を処理する方法であって、
複数の可能性のある合致を検出するためにフィルタ回路にデータサブ文字列について問い合わせをするステップであって、各可能性のある合致がデータサブ文字列とパターン断片との間の可能性のある合致を表し、複数のパターンの各パターンが複数の対応するパターン断片を含み、フィルタ回路がパターン断片でプログラムされている、問い合わせをするステップと、
複数のパターンセットを決定するために、検出された可能性のある合致を削減ステージに適用するステップであって、各パターンセットが、検出された可能性のある合致と対応し、（１）対応する可能性のある合致を生じたパターン断片に対応するパターンを表すデータおよび（２）該パターンに関連する許容可能な誤りを表すデータを含む、適用するステップと、
配信されたパターンセット内の許容可能な誤りデータを考慮して、配信されたパターンセット内のパターンのいずれかと、データ文字列の少なくとも一部が近似合致するか否かを決定するために、決定されたパターンセットを近似合致エンジンに配信するステップとを含む、方法。
【請求項２】

フィルタ回路がブルームフィルタ回路である、請求項１に記載の方法。
【請求項３】
配信されたパターンセット内の許容可能な誤りデータを考慮して、データ文字列部分と配信されたパターンセット内のパターンとの間で何らかの近似合致が存在するか否かを近似合致エンジンを用いて判定するステップをさらに含む、請求項１または２に記載の方法。
【請求項４】
少なくとも複数のパターンであるパターンを複数のパターン断片にスライスするステップと、
パターン断片を用いてブルームフィルタ回路をプログラムするステップとをさらに含む、請求項２に記載の方法。
【請求項５】
再構成可能なハードウェアを用いて、判定するステップを実行するステップをさらに含む、請求項３に記載の方法。
【請求項６】
問い合わせするステップおよび適用するステップをパイプライン方式で実行するステップをさらに含む、請求項１から５のいずれか一項に記載の方法。
【請求項７】
再構成可能なハードウェアを用いて、問い合わせするステップおよび適用するステップを実行するステップをさらに含む、請求項１から６のいずれか一項に記載の方法。
【請求項８】
ブルームフィルタ回路が複数の並列ブルームフィルタ回路を含み、並列ブルームフィルタ回路の各々が、各並列ブルームフィルタ回路についてのパターン断片長が他の並列ブルームフィルタ回路についての他のパターン断片長と異なり、問い合わせステップが、複数のデータサブ文字列を並列ブルームフィルタ回路に並列に同時に提供するステップを含む、請求項４に記載の方法。
【請求項９】
ブルームフィルタ回路が、データ文字列の少なくとも一部を記憶するシフトレジスタをさらに含み、方法が、シフトレジスタを通してデータ文字列のデータシンボルを流すステップと、問い合わせするステップのためにシフトレジスタからデータサブ文字列を読むステップとをさらに含む、請求項２、４又は８のいずれか一項に記載の方法。
【請求項１０】
各並列ブルームフィルタ回路が、複数のビットベクトルを記憶するように構成されており、各ビットベクトルが複数のビットを含み、各ビットが値を有し、問い合わせするステップが、
少なくとも１つのハッシュキーを生成するために、各データサブ文字列を少なくとも１つのハッシュ関数に適用するステップと、
少なくとも１つの生成されたハッシュキーに基づいて、並列ブルームフィルタ回路からビットベクトルを取り出すステップと、
少なくとも１つの生成されたハッシュキーに基づいて、取り出されたビットベクトルの複数のビットの位置を選択するステップと、
選択されたビットの位置にあるビットの値に基づいて、該データサブ文字列とパターン断片との間に可能性のある合致が存在するか否かを判定するステップとをさらに含む、請求項８に記載の方法。
【請求項１１】
各データサブ文字列を少なくとも１つのハッシュ関数に適用するステップが、複数のハッシュキーを生成するために複数のハッシュ関数に各データサブ文字列を適用するステップを含み、ビットベクトルを取り出すステップが、生成されたハッシュキーに基づいて並列ブルームフィルタ回路からビットベクトルを取り出すステップを含み、ビットの位置を選択するステップが、複数の他の生成されたハッシュキーに基づいて、取り出されたビッ

トベクトルの複数のビットの位置を選択するステップを含む、請求項１０に記載の方法。
【請求項１２】
各データサブ文字列を少なくとも１つのハッシュ関数に適用するステップが、単一のハッシュキーを生成するために単一のハッシュ関数に各データサブ文字列を適用するステップを含み、ビットベクトルを取り出すステップが、生成されたハッシュキーの一部に基づいて並列ブルームフィルタ回路からビットベクトルを取り出すステップを含み、ビットの位置を選択するステップが、生成されたハッシュキーの別の部分に基づいて、取り出されたビットベクトルの複数のビットの位置を選択するステップを含む、請求項１０に記載の方法。
【請求項１３】
削減ステージが、（１）各パターン断片識別子が問い合わせするステップの結果として生成されたデータによってインデックス付けされる、複数のパターン断片識別子を記憶し、（２）各パターン識別子がパターン断片識別子によってインデックス付けされる、複数のパターン識別子を記憶し、（３）各パターンセット対がパターン識別子によってインデックス付けされる、複数のパターンセット対を記憶するように構成されており、各パターンセット対が、（ａ）パターンを表すデータおよび（ｂ）該パターンに関連する許容可能な誤りを表すデータを含み、可能性のある合致を適用するステップが、
問い合わせするステップの結果として生じたデータに基づいて、記憶されているパターン断片識別子の少なくとも１つを取り出すステップと、
各取り出されたパターン断片識別子に基づいて、記憶されているパターン識別子の少なくとも１つを取り出すステップと、
各取り出されたパターン識別子に基づいて、記憶されているパターンセット対の少なくとも１つを取り出すステップとを含み、取り出された少なくとも１つのパターンセット対が、パターンセットの少なくとも一部を規定する、請求項１から１２のいずれか一項に記載の方法。
【請求項１４】
パターン断片識別子を取り出すステップが、少なくとも１つの生成されたハッシュキーに対応するデータに基づいて、記憶されているパターン断片識別子の少なくとも１つを取り出すステップを含む、請求項１３に記載の方法。
【請求項１５】
削減ステージが、（１）各パターン識別子がパターン断片によってインデックス付けされる、複数のパターン識別子を記憶し、（２）各パターンセット対がパターン識別子によってインデックス付けされる、複数のパターンセット対を記憶するように構成されており、各パターンセット対が、（ａ）パターンを表すデータおよび（ｂ）該パターンに関連する許容可能な誤りを表すデータを含み、可能性のある合致を適用するステップが、
可能性のある合致を生じたデータサブ文字列に基づいて、記憶されているパターン識別子の少なくとも１つを取り出すステップと、
各取り出されたパターン識別子に基づいて、記憶されているパターンセット対の少なくとも１つを取り出すステップとを含み、取り出された少なくとも１つのパターンセット対が、パターンセットの少なくとも一部を規定する、請求項１から１２のいずれか一項に記載の方法。
【請求項１６】
複数のデータシンボルを含むデータ文字列が複数のパターンのうちのいずれかとの近似合致を含むか否かの決定を容易化するために、データ文字列内の複数のデータサブ文字列を処理するシステムであって、
複数の可能性のある合致を検出するためにデータサブ文字列によって問い合わせされるように構成されているフィルタ回路であって、各可能性のある合致がデータサブ文字列とパターン断片との間の可能性のある合致を表し、複数のパターンの各パターンが複数の対応するパターン断片を含み、フィルタ回路がパターン断片でプログラムされている、フィルタ回路と、

フィルタ回路と通信する削減ステージであって、削減ステージが、（１）検出された可能性のある合致を処理し、（２）処理に応じて複数のパターンセットを決定し、各パターンセットが可能性のある合致に対応して（ａ）対応する可能性のある合致を生じたパターン断片に対応するパターンを表すデータと、（ｂ）該パターンに関連する許容可能な誤りを表すデータとを含み、（３）決定されたパターンセット内の許容可能な誤りデータを考慮して、データ文字列の少なくとも一部が、決定されたパターンセット内のいずれかパターンと近似合致するか否かを判定するために、近似合致エンジンに決定されたパターンセットを出力するように構成されている、削減ステージとを含む、システム。

【請求項１７】

フィルタ回路がブルームフィルタ回路である、請求項１６に記載のシステム。

【請求項１８】

削減ステージと通信しており、決定されたパターンセット内の許容可能な誤りデータを考慮して、データ文字列部分と決定されたパターンセット内のパターンとの間で何らかの近似合致が存在するか否かを判定するように構成されている近似合致エンジンをさらに含む、請求項１６または１７に記載のシステム。

【請求項１９】

ブルームフィルタ回路および削減ステージがパイプライン方式で動作するように構成されている、請求項１７に記載のシステム。

【請求項２０】

ブルームフィルタ回路および削減ステージが再構成可能なハードウェアを用いて実装されている、請求項１７または１９に記載のシステム。

【請求項２１】

近似合致エンジンが再構成可能なハードウェアを用いて実装されている、請求項１８に記載のシステム。

【請求項２２】

パターン断片が、異なる長さの複数のパターン断片である、請求項１６から２１のいずれか一項に記載のシステム。

【請求項２３】

同じパターンに対応するパターン断片が互いに重なり合わない、請求項１６から２１のいずれか一項に記載のシステム。

【請求項２４】

ブルームフィルタ回路が複数の並列ブルームフィルタ回路を含み、各並列ブルームフィルタ回路についてのパターン断片長が他の並列ブルームフィルタ回路についての他のパターン断片長と異なり、ブルームフィルタ回路が、複数のデータサブ文字列を並列ブルームフィルタ回路に並列に同時に提供するようにさらに構成されている、請求項１７、１９または２０のいずれか一項に記載のシステム。

【請求項２５】

ブルームフィルタ回路が、データ文字列の少なくとも一部を記憶するシフトレジスタをさらに含み、ブルームフィルタ回路が、（１）シフトレジスタを通してデータ文字列のデータシンボルを流し、（２）並列ブルームフィルタ回路への配信のためにシフトレジスタからデータサブ文字列を読むようにさらに構成されている、請求項２４に記載のシステム。

【請求項２６】

各並列ブルームフィルタ回路が、複数のビットベクトルを記憶するように構成されており、各ビットベクトルが複数のビットを含み、各ビットが値を有し、ブルームフィルタ回路が、（１）少なくとも１つのハッシュキーを生成するために、各データサブ文字列を少なくとも１つのハッシュ関数に適用し、（２）少なくとも１つの生成されたハッシュキーに基づいて、並列ブルームフィルタ回路からビットベクトルを取り出し、（３）少なくとも１つの生成されたハッシュキーに基づいて、取り出されたビットベクトルの複数のビットの位置を選択し、（４）選択されたビットの位置にあるビットの値に基づいて、該デー

タサブ文字列とパターン断片との間に可能性のある合致が存在するか否かを判定するようにさらに構成されている、請求項２４に記載のシステム。
【請求項２７】
各並列ブルームフィルタ回路が、（１）複数のハッシュキーを生成するために複数のハッシュ関数に各データサブ文字列を適用し、（２）生成されたハッシュキーに基づいて並列ブルームフィルタ回路からビットベクトルを取り出し、（３）複数の他の生成されたハッシュキーに基づいて、取り出されたビットベクトルの複数のビットの位置を選択するようにさらに構成されている、請求項２６に記載のシステム。
【請求項２８】
各並列ブルームフィルタ回路が、（１）単一のハッシュキーを生成するために単一のハッシュ関数に各データサブ文字列を適用し、（２）生成されたハッシュキーの一部に基づいて並列ブルームフィルタ回路からビットベクトルを取り出し、（３）生成されたハッシュキーの別の部分に基づいて、取り出されたビットベクトルの複数のビットの位置を選択するようにさらに構成されている、請求項２６に記載のシステム。
【請求項２９】
削減ステージが第１のテーブル、第２のテーブルおよび第３のテーブルを含み、第１のテーブルは、各パターン断片識別子がブルームフィルタ回路によって生成されたデータによってインデックス付けされる、複数のパターン断片識別子を記憶するように構成されており、第２のテーブルは、各パターン識別子がパターン断片識別子によってインデックス付けされる、複数のパターン識別子を記憶するように構成されており、第３のテーブルは、各パターンセット対がパターン識別子によってインデックス付けされる、複数のパターンセット対を記憶するようにさらに構成されており、各パターンセット対が、（ａ）パターンを表すデータおよび（ｂ）該パターンに関連する許容可能な誤りを表すデータを含み、削減ステージが、（１）ブルームフィルタ回路によって生じたデータに基づいて、第１のテーブルから記憶されているパターン断片識別子の少なくとも１つを取り出し、（２）各取り出されたパターン断片識別子に基づいて、第２のテーブルから記憶されているパターン識別子の少なくとも１つを取り出し、（３）各取り出されたパターン識別子に基づいて、第３のテーブルから記憶されているパターンセット対の少なくとも１つを取り出すようにさらに構成されており、少なくとも１つの取り出されたパターンセット対が、パターンセットの少なくとも一部を規定する、請求項１６から２８のいずれか一項に記載のシステム。
【請求項３０】
削減ステージが第１のデータ構造および第２のデータ構造を含み、第１のデータ構造は、各パターン識別子がパターン断片によってインデックス付けされる、複数のパターン識別子を記憶するように構成されており、第２のデータ構造は、各パターンセット対がパターン識別子によってインデックス付けされる、複数のパターンセット対を記憶するように構成されており、各パターンセット対が、（ａ）パターンを表すデータおよび（ｂ）該パターンに関連する許容可能な誤りを表すデータを含み、削減ステージが、（１）可能性のある合致を生じたデータサブ文字列に基づいて、第１のデータ構造から記憶されているパターン識別子の少なくとも１つを取り出し、（２）各取り出されたパターン識別子に基づいて、第２のデータ構造から記憶されているパターンセット対の少なくとも１つを取り出すようにさらに構成されており、少なくとも１つの取り出されたパターンセット対が、パターンセットの少なくとも一部を規定する、請求項１６から２８のいずれか一項に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、パターンの大きなセットを用いた近似パターン合致の分野に関する。特に、本発明は、多数のグループのパターンを用いた近似パターン合致のためのスケーラブルフィルタ回路および削減ステージに関する。

【背景技術】
【０００２】
近似パターンまたは文字列合致は、多数の重要なアプリケーションで生じる重要な問題である。限定ではないが、これらは、コンピュータ利用生命工学、データベースおよびコンピュータ通信を含み得る。このタスクは、通常、特定の数の誤りを許しながらの、特定のパターンまたはパターンセットの間での合致の検索を含む。一例として、２つの誤りを許しながら、単語「ｑｕｅｕｉｎｇ」を検索することを望むことが可能である。これは、１つの文字が挿入されている単語「ｑｕｅｕｅｉｎｇ」、および１つの文字が置換されて１つの文字が削除されている「ｃｕｅｉｎｇ」等の結果を返す可能性がある。特定の数の誤りを許可することによって、これは、通常のつづりの変化または誤りを捕捉して検索し、それでも所望のパターンの発見を可能にする。近似パターン合致は、複雑なタスクであるだけではなく、途方もない量のコンピュータリソースを必要とする。
【０００３】
通常、高速のフィルタリングステップがあり、次に全部の近似合致機能を実行する検証ステップが続く。この従来技術のフィルタ技術の一例は、図１を参照すると示されており、図番１０で概略的に示されている。この通常の手法は、図番１２で示されるようなパターン「Ｐ」を、重複しない一連のサブパターンであるｋ＋１個のパターン断片にスライスし、テキストとパターン断片との間の正確な合致を検索する。この場合、「ｋ」は、最大編集距離ｅｄ（$T_{i...j}$，Ｐ）であり、この限定でない例では図番１４で示されるように数字２（２）で示されている、許容可能な誤りの数と等しい。
【０００４】
データ文字列$T_{i...j}$１６は、次に、データ文字列１６のうちの少なくとも１つのサブ文字列が、パターン「Ｐ」１２と関連する重複しないサブパターンのうちの少なくとも１つと合致する出現に関して解析される。この手法は、以下の特性に依存する：
ａ．文字列Ｓ＝$T_{a...b}$が最大ｋの誤りでパターンＰと合致し、Ｐ＝$p_1...p_j$（重複しないサブパターンの配列）である場合、Ｓのいくつかのサブ文字列は複数のp_iのうちの少なくとも１つと最大
【数１】

$$\lfloor k/j \rfloor$$

の誤りで合致する。
ｂ．ｅｄ（$T_{i...j}$，Ｐ）≦ｋである文字位置ｉ≦ｊが存在する場合、$T_{j-m+1...j}$は、Ｐのうちの少なくともｍ－ｋ個の文字を含み、ここでｍはパターンの（文字の）大きさである。
ｃ．それゆえに、Ｐをｋ＋１個の断片（重複しないサブパターン）にスライスする場合、断片のうちの少なくとも１つが正確に合致するはずである。
【０００５】
それゆえに、「Ｐ」１２を、誤りの総数「ｋ」１４に一個（１）加えた数の重複しないサブパターン断片にスライスする場合、重複しないサブパターン断片のうちの少なくとも１つが正確に合致するはずである。図１の例に示されるように、データ文字列$T_{i...j}$１６が、重複しないサブパターンのｋ＋１または三個（３）の断片に分割されている。それゆえに、この三個（３）の断片は、図番１８で示される「ａｂｒａ」、図番２０で示される「ｃａｄａ」、および図番２２で示される「ｂｒａ」である。この例では、文字「ｂｒ」が置換され、文字「ｂ」が削除されている２つの誤りと共に、図番２０で示される「ｃａｄａ」が正確に合致している。
【発明の開示】
【発明が解決しようとする課題】
【０００６】
可能性のある合致パターンの大きなセットを用いた、大量の入力データを利用するパタ

ーン合致に関する高速で費用効率が高い機構に対する著しい必要性が存在する。

【課題を解決するための手段】

【０００７】

本発明の一態様では、少なくとも１つのサーチエンジンを用いて、各々のパターンが所定の許容可能な誤りを有する１つまたは複数のパターンと合致するデータセグメントについてデータストリームを検査する方法が開示されている。この方法は、各々のパターンが関連する許容可能な誤りを有する１つまたは複数のパターンを検出するように各々の並列フィルタ機構が構成されている複数の並列フィルタ機構を用いて、複数のパターンのシンボルの組み合わせに関してデータストリームをフィルタリングすることと、複数の並列フィルタ機構を用いて複数の可能性のあるパターン断片の合致を検出することと、複数の並列フィルタ機構から、各々の合致パターンが関連する許容可能な誤りを有する、複数の可能性のある合致パターンを特定することと、削減ステージを用いて、特定された複数の可能性のある合致パターンを、各々の合致パターンが関連する許容可能な誤りを有する、可能性のある合致パターンのセットへと削減することと、関連するデータ、および各々の合致パターンが関連する許容可能な誤りを有する、可能性のある合致パターンの削減されたセットを検証ステージに提供することと、関連するデータおよび可能性のある合致パターンの削減されたセットを利用する近似合致エンジンを含む検証ステージを用いて、複数のパターンのシンボルの組み合わせおよび関連する許容可能な誤りからデータストリームでのパターン合致の存在を検証することとを含む。

【０００８】

本発明の別の態様では、少なくとも１つのサーチエンジンを用いて、各々のパターンが所定の許容可能な誤りを有する１つまたは複数のパターンと合致するデータセグメントについてデータストリームを検査する方法が開示されている。この方法は、並列ブルーム（Ｂｌｏｏｍ）フィルタのセット、並列ブルームフィルタ配列のセットまたは単一のハッシュキージェネレータを利用する並列ブルームフィルタ配列のセットから成るグループであり、各々のパターンが関連する許容可能な誤りを有する１つまたは複数のパターンを検出するように各々の並列フィルタ機構が構成されている、複数の並列フィルタ機構を用いて、複数のパターンのシンボルの組み合わせに関してデータストリームをフィルタリングすることと、複数の並列フィルタ機構を用いて複数の可能性のあるパターン断片の合致を検出することと、複数の並列フィルタ機構から、各々の合致パターンが関連する許容可能な誤りを有する、複数の可能性のある合致パターンを特定することと、特定された複数の可能性のある合致パターンを、各々の合致パターンが関連する許容可能な誤りを有する、可能性のある合致パターンのセットへと削減することと、関連するデータ、および各々の合致パターンが関連する許容可能な誤りを有する、可能性のある合致パターンの削減されたセットを検証ステージに提供することと、関連するデータおよび可能性のある合致パターンの削減されたセットを利用する近似合致エンジンを含む検証ステージを用いて、複数のパターンのシンボルの組み合わせおよび関連する許容可能な誤りからデータストリームでのパターン合致の存在を検証することとを含む。

【０００９】

本発明のさらに別の態様では、少なくとも１つのサーチエンジンを用いて、各々のパターンが所定の許容可能な誤りを有する１つまたは複数のパターンと合致するデータセグメントについてデータストリームを検査する方法およびシステムが開示されている。この方法は、少なくとも１つの検索エンジンを用いて、誤検出誤りを有する１つまたは複数のパターン断片と合致するデータセグメントについてデータストリームを検査するために、単一のハッシュ値から複数のハッシュ値を抽出するために単一のハッシュキージェネレータを利用することと、複数の並列ブルームフィルタ配列を用いて複数のハッシュ値を利用することとを含む。

【００１０】

本発明のさらに別の態様では、少なくとも１つのサーチエンジンを用いて、各々のパターンが所定の許容可能な誤りを有する１つまたは複数のパターンと合致するデータセグメ

ントについてデータストリームを検査するシステムが開示されている。このシステムは、各々のパターンが関連する許容可能な誤りを有する１つまたは複数のパターンを検出するように各々の並列フィルタ機構が構成されている複数の並列フィルタ機構を利用し、複数のパターンのシンボルの組み合わせに関してデータストリームをフィルタリングし、複数の可能性のあるパターン断片の合致を検出し、各々の合致パターンが関連する許容可能な誤りを有する、複数の可能性のある合致パターンを特定するフィルタステージと、特定された複数の可能性のある合致パターンを、各々の合致パターンが関連する許容可能な誤りを有する、可能性のある合致パターンのセットへと削減する削減ステージと、複数のパターンのシンボルの組み合わせからデータストリームでのパターン合致の存在を検証するために、関連するデータおよび可能性のある合致パターンの削減されたセットおよび関連する許容可能な誤りを受信および利用する、近似合致エンジンを含む検証ステージとを含む。

【００１１】

本発明のさらに別の態様では、少なくとも１つのサーチエンジンを用いて、各々のパターンが所定の許容可能な誤りを有する１つまたは複数のパターンと合致するデータセグメントについてデータストリームを検査するシステムが開示されている。このシステムは、並列ブルームフィルタのセット、並列ブルームフィルタ配列のセットまたは単一のハッシュキージェネレータを利用する並列ブルームフィルタ配列のセットから成るグループであり、各々のパターンが関連する許容可能な誤りを有する１つまたは複数のパターンを検出するように各々の並列フィルタ機構が構成され、複数のパターンのシンボルの組み合わせに関してデータストリームをフィルタリングし、複数の可能性のあるパターン断片の合致を検出し、各々の合致パターンが関連する許容可能な誤りを有する、複数の可能性のある合致パターンを特定する複数の並列フィルタ機構と、特定された複数の可能性のある合致パターンを、各々の合致パターンが関連する許容可能な誤りを有する、可能性のある合致パターンのセットへと削減する削減ステージと、複数のパターンのシンボルの組み合わせからデータストリームでのパターン合致の存在を検証するために、関連するデータおよび可能性のある合致パターンの削減されたセットおよび関連する許容可能な誤りを受信および利用する、近似合致エンジンを含む検証ステージとを含む。

【００１２】

例示的であり、限定ではないが、本発明の可能性のあるアプリケーションの例は、コンピュータ通信ネットワークに関する侵入検知システム（ＩＤＳ）、コンピュータ計算による生物学および遺伝学、構造化および非構造化テキストに対するテキスト検索、および光学文字走査（ＯＣＳ）からのテキスト検索を含む。

【００１３】

本発明の付加的な態様は、限定ではないが、各パターンが、例えば何万パターン以上の多数のパターン断片を含み得るその許容可能な誤りを特定し得る、複数のパターンを用いた近似合致に関するフィルタリング技術と、並列合致動作を実行するために、および多種多様の（パターン長、許容可能な誤り）組み合わせをサポートするために、１つが各パターン断片長のためのものである、正確な合致エンジンの並列セットを利用することと、各パターンに特定の数の誤りを許容することと、並列ハードウェア検索実装に対する従順さおよびそのような実装が高速の検索結果を提供し得ることと、テキストの範囲に関して可能性のある合致パターンの数を限定することによって検証ステージを単純化し、検索エンジンが付加的な可能性のある検索結果をより短い周期で処理することを可能にして、全システムが非常に高速で動作しながら能力を向上させることと、各正確な合致エンジンに対してブルームフィルタ配列を利用することと、１つだけのハッシュ関数ジェネレータを使用することによって各ブルームフィルタ配列を効果的に実装することとを含む。

【００１４】

本発明のさらに別の態様は、検証ステージでの検索の範囲が可能性のある合致パターンのより小さいセットへと削減される削減ステージである。これらの技術は、各パターンおよびその許容可能な誤りが一度だけ記憶されることを可能にする、断片およびパターンの

間の間接層を使用する。データ構造を単純化し、それらをハードウェア実装に従順にする第１の例示的な技術が存在する。この技術は、可能性のある合致断片に関する断片識別子を取り出すために、ビンインデックスを使用する参照を含む。第２の参照は、断片を含むパターンに関するパターン識別子を取り出すために断片識別子を使用する。第３の参照は、検証エンジンによって考慮されるパターンおよび関連する許容可能な誤りの対を取り出すためにパターン識別子を使用する。断片を含むパターンに関するパターン識別子を解決するために、正確な合致エンジンでの合致を生じたテキスト断片を使用する第２の例示的であるが、限定ではない技術が存在する。パターン識別子は、検索エンジンによって考慮されるパターンおよび関連する許容可能な誤りの対を取り出すために使用される。

【００１５】

これらは、本発明の無数の態様のうちのただ一部であり、本発明に関連付けられている無数の態様の全てを含むリストとみなされるべきではない。

【００１６】

本発明をより理解するために、添付の図面を参照されたい。

【発明を実施するための最良の形態】

【００１７】

以下の詳細な説明では、本発明の完全な理解をもたらすために多数の特定の詳細が示されている。しかしながら、当業者は、本発明がこれらの特定の詳細なしで実施され得ることを理解するであろう。別の例では、よく知られている方法、手順および構成要素は、本発明を分かりにくくさせるので詳細には説明されていない。

【００１８】

本発明は、パターンの大きなセットを用いた近似パターン合致のためのスケーラブルなフィルタ回路である。フィルタ回路は、記憶されているパターンのセットの間で可能性のある合致を検証し、各パターンが許容可能な誤りの数および文字の入力ストリームを特定する。許容可能な誤りの数は、所定のものであるか、または単一文字の付加、削除および置換の数を数える一般的な編集距離尺度を使用して特定される。可能性のある合致が検出されると、合致パターンまたは複数の可能性のある合致パターン同様に、入力データストリームでの単数または複数の位置が特定される。本発明は、可能性のある合致パターンの総数から、前に特定された可能性のある合致パターン（単数または複数）を利用し、入力データストリームのデータセグメント（単数または複数）での近似合致を検索する検証ステージと協働して動作するように設計されている。

【００１９】

単一パターンに関する検索の方法論は、以下のように規定され得る：「ｋ」の許容可能な誤りでテキスト「Ｔ」でのパターン「Ｐ」のインスタンスを特定する、ここで「ｋ」は最大編集距離である。編集距離は、全ての誤りが、必ずではないが通常は、同じ重み付けを共有する、単一文字の挿入、削除および置換の数として規定されている。一般に、転置等の別の種類の誤りが距離尺度に含まれてもよく、各種類の誤りは固有の重み付けが割り当てられ得る。

【００２０】

「ｍ」がパターンの文字としての大きさであり、「ｎ」がテキストの文字としての大きさであると仮定すると、誤りレベルは、特定のパターンに対して、許容可能誤りの数とパターンの大きさの比率、即ち「α＝ｋ／ｍ」として規定され得る。誤りレベルならびにパターンおよびテキストが構成されるアルファベットの大きさ「σ」は、合致が見つかる確率に影響を及ぼす。無作為に構成されているテキストおよび無作為に構成されているパターンを仮定しての合致確率の式「ｆ（ｍ，ｋ）」は、以下の式：

【数２】

$$f(m,k)=\left(\frac{e^2}{\sigma(1-\alpha)^2}\right)^{m(1-\alpha)}$$

で提供される。ここで「ｅ」は自然対数の底である。

【００２１】

フィルタリングアルゴリズムは種々の近似パターン合致問題に対してより良い成績を達成すると考えられている。一般的な手法は、可能性のある合致を特定するためにテキストの小セクションに対し単純な検索を実行するものである。可能性のある合致が見つかると、テキストのその範囲は、それが実際に特定パターンに対する近似合致であるか否かを確かめるために検査される。一般に、検証は、あらゆる近似パターン合致アルゴリズムで実行され、多かれ少なかれフィルタリング動作に結び付いている可能性がある。

【００２２】

本発明のシステムの概略的な模式図が図２に示されており、図番３０によって示されている。高いデータ転送速度でフィルタ回路に供給可能な種々の他のデータソースと同様に、通信リンク、ディスク、レイド（ｒｅｄｕｎｄａｎｔ　ａｒｒａｙ　ｏｆ　ｉｎｄｅｐｅｎｄｅｎｔ　ｄｉｓｋｓ，ＲＡＩＤ）またはストレージエリアネットワーク（ＳＡＮ）のいずれかを介して大量のデータが入力として提供され得る。このデータ入力は図番３２で示され、図番３４で示されるネットワーク入力を介して同様に提供され得る。高速データは、図番３６で示される高帯域相互接続を介して高速で提供され得る。この高速データは、可能性のある合致に対して入力データを入力パターンのセットに関してスキャンするフィルタ回路３８を次に通される。次に、フィルタ回路３８で合致を取り出すデータセグメントを処理する際に、検証ステージ４２によって考慮される必要がある可能性のある合致パターンのセットを削減する削減ステージ４０がフィルタ回路３８と検証ステージ４２との間に存在する。検証ステージ４２は、入力パターン４６のセットに対し、合致が存在するか否かを検証するための全ての適切な動作を実行する。次に、検索結果が、図番４４で示されるように提供される。所定の入力パターン４６がフィルタ回路３８、削減ステージ４０および検証ステージ４２に提供される。

【００２３】

テキストの特定の窓では、正確な合致を検索することが可能であり、パターン断片のいずれもが所定数「ｒ」のパターンによって特定される。特定のパターン「ｉ」が「k_i」個の誤りを許容する場合、パターン断片の総数は式：

【数３】

$$p = \sum_{i=1}^{r} (k_i + 1)$$

によって示される。

【００２４】

このフィルタリング手法は、ブルームフィルタの並列セット、並列ブルームフィルタ配列のセット、または単一ハッシュ関数ジェネレータを利用するブルームフィルタ配列のセットを含み得る、並列フィルタ（ｐａｒａｌｌｅｌ　ｆｉｌｔｅｒ）機構と共に利用される。本発明の基本的なハードウェア実装の概略図である図３で示されるように、本発明が図番５０で示されており、図番５４で示される多数のブルームフィルタを含む。従来のブルームフィルタ５４では、「ｂ」個のハッシュ値を使用してセットに要素が挿入され、要素がキーとして利用され、そこで各ハッシュ値がＢ－ビットベクトルでのビット位置を特定する。ｂビット位置の各々でのビットは１に設定されるのが好ましい。ビットが既に１に設定されている場合、変更はされない。特定の要素が、ブルームフィルタ５４によって示されるセットの構成要素であるか否かを検査するために、要素および同じｂハッシュ関数がｂハッシュ値を計算するのに利用される。ベクトルの全てのｂビットが１に設定されている場合、その要素がセットの構成要素であることが宣言される。

【００２５】

ブルームフィルタ５４は、検出漏れを生じない。要素がセットの構成要素である場合、要素がセットに挿入される際にＢ－ビットベクトルのｂビット位置が１に設定される。付

加的な要素のセットへの挿入は、ベクトル内のいずれのビットをもリセットしない。しかしながら、ブルームフィルタ５４は、所定の確率で検出漏れを生じる。この確率は、式：
【数４】

$$f=\left(1-e^{\frac{-pb}{B}}\right)^{b}$$

で計算され得る。以下の関係が成り立つ場合：
【数５】

$$b=\frac{B}{p}\ln 2 \text{の場合}: f=(1/2)^{b}$$

【００２６】
複数のパターンでの近似合致フィルタリングと、各パターンがその許容可能な誤りを特定することとは、種々の長さのパターン断片のセットを生じる。好ましくは、必ずではないが、ブルームフィルタ５４は、可能性のある各パターン断片長に対して１つのブルームフィルタ回路５４の固定長要素を記憶している。それゆえに、可能性のあるパターン断片長の範囲は、ある範囲内に制限される。
【００２７】
l_{min}が最小パターン断片長である場合、l_{min}は、パターンｍの大きさを最大編集距離ｋに１を加えたもので割った値より小さいか等しい：
【数６】

$$l_{min}\leq\left\lfloor\frac{m}{k+1}\right\rfloor$$

【００２８】
l_{max}が最大断片長である場合、l_{max}は、パターンｍの大きさを最大編集距離ｋに１を加えたもので割った値より大きいか等しい：
【数７】

$$\left\lceil\frac{m}{k+1}\right\rceil\leq l_{max}$$

【００２９】
ブルームフィルタ５４の各ブルームフィルタがパターン断片長に対応する際に必要であるブルームフィルタ５４の総数は：
$l_{max}-l_{min}+1$
である。
【００３０】
好ましい手法は、図３の図番５０で概略的に示されているように、ブルームフィルタ５４の各々に並行して問い合わせることである。種々のストライドのテキスト窓が各ブルームフィルタ５４への入力キーとして選択可能であるように、ブルームフィルタ５４の各々は、パターン断片長に対応している。
【００３１】
ブルームフィルタ５４のいずれかが検出の合致５６を生じると、データのセグメントまたはテキスト窓５８、入力ストリームのデータのセグメントの位置５９および付加的な合致メタデータ６０が削減ステージ４０に送信される。削減ステージ４０で利用される技術は、可能性のある合致パターンを著しく、例えば、１００００超を１０未満へと、削減することが可能である。
【００３２】

削減ステージ４０から渡された結果は、近似合致検索エンジンを含む検証ステージ４２に進む。検証ステージ４２によって考慮される候補パターンの数を削減することによって、検証ステージが所与の時間でより多くの可能性のある検索結果を処理することが可能になり、従って、高速で動作しながら全システムの能力面を向上させることが可能になる。

【００３３】

通常、ブルームフィルタ配列５４は、セットメンバーシップ問い合わせに必要な、例えば、ランダムアクセスメモリ（ＲＡＭ）等のメモリアクセスの数を最小化する。さらに、ブルームフィルタ配列５４は、Ｂ－ビットベクトルを大きさ「ｑ＝Ｂ／Ｗ」の「Ｗ」個のベクトルに分割し、ここで「ｑ」はメモリのワードサイズである。好ましくは、プリフィルタハッシュ関数を使用して、記憶されている要素が「Ｗ」個のベクトル（メモリワード）上に均等に分配されていることが好ましい可能性がある。これは、「Ｗ」個の「ｑ」－ビットブルームフィルタの配列を生じる。

【００３４】

プログラミングの際に「ｑ」－ビットブルームフィルタ５４のビットが設定され、次に、問い合わせの際に「ｂ」個のハッシュ関数を使用して検査される。ブルームフィルタ配列５４への問い合わせは、１個のメモリが「ｑ」－ビットベクトルを読んで取ってくることを必要とする。図４で示されて図番７０で概略的に示されるように、レジスタ８０およびビット選択回路８２を使用することによって、「ｂ」個のハッシュ関数によって特定されるビット位置の検査がオンチップでパイプライン方式で実行され得る。

【００３５】

このアプリケーションでは、キー（即ち、パターン断片）が図番７２によって示されている。このキーは、「ｗ」個のベクトルのリストの特定の「ｑ」－ビットベクトルを特定するために、プレフィルタハッシュ関数７３によって使用される。「ｗ」個のベクトルのリストの特定の「ｑ」－ビットベクトルは、例えば、ＲＡＭ等のメモリの列７４で示されており、そこで具体的で例示的なベクトルが図番７６で特定されている。これらの問い合わせは、レジスタ８０内でのビット位置を特定する、図番７８で示される一連の「ｂ」個のハッシュ関数で検査され、次に、合致検出機能８２に提供される。「ｂ」個のビット位置の全てが１に設定されている場合、このキーは可能性のある合致パターンに関するパターン断片である。

【００３６】

図５に示されるように、ブルームフィルタ配列５４を実装するために必要なロジックの量が最小化され得るのが好ましい。このロジックは、図番９０で概略的に示されている。図番９２で示される単一のハッシュ関数が存在する。単一の乱数値の生成が存在する。この乱数値からのビットのサブセットがプレフィルタハッシュアドレスおよび図番９４で示される「ｂ」個のフィルタビット位置を構成するために利用される。「ｗ」個のベクトルのリストの特定の「ｑ」－ビットベクトルが、例えば、ＲＡＭ等のメモリの列９６で示されており、そこで具体的で例示的なベクトルが図番９８で特定されている。この特定のベクトル９８は、レジスタ１００に渡され、次に、合致パターン検出機能１０２に渡される。このビット選択値９４は、サイズが少なくともlog_2（Ｗ）＋（ｂ＊log_2（ｑ））ビットである必要がある。ハッシュ関数９２のH_3クラスは、このアプリケーションに充分に広範な値を提供し得る例示的で限定ではないハッシュ関数例である。

【００３７】

利用可能な再構成可能なハードウェアの例示的で限定ではない例は、Ｘｉｌｉｎｘ（登録商標）ＶｉｒｔｅｘＩＩ（登録商標）４０００シリーズＦＰＧＡを含むＦＰＧＡｓ、即ち、フィールドプログラマブルゲートアレイを含む。Ｘｉｌｉｎｘ社は、９５１２４－３４００、カリフォルニア州、サンノゼ、ロジックドライブ２１００に事業所を有するデラウェア州の会社である。ＶｉｒｔｅｘＩＩ（登録商標）のデバイスシリーズの内蔵メモリの例示的で限定ではない例は、１２０個の１８キロバイトブロックのランダムアクセスメモリ（ブロックＲＡＭｓ）を含む。これらのブロックＲＡＭは、例示的で限定ではないが、最大ワード長３６ビット×５１２ワードを用いて種々の大きさのワードに構成され得る

。

【００３８】

誤検出の確率に関する前式を利用し、一定のハッシュ性能を仮定すると、１８キロバイトブロックＲＡＭを用いて実装されているブルームフィルタ配列５４は、ｂビット位置の数が４と等しい場合、３，１９４個の要素のセットの誤りの確率を０．０６３と示し得る。ｂビット位置の数が３と等しい場合、能力は４，２５９個の要素に増加するが、誤検出の可能性が０．１２５に増大する。

【００３９】

前述したように、１つのブルームフィルタ配列５４が各固有のパターン断片長に対して必要である。さらに、データ入力速度の歩調をとる必要がある並列回路の数の検討材料も存在する。例示的で限定ではない例では、周期（６４－ビットインターフェース）当たり８個の新規のＡＳＣＩＩ文字を受け入れるシステムは、並列に動作する回路の８個のインスタンスを必要とする。ＶｉｒｔｅｘＩＩ４０００（登録商標）の場合、各回路インスタンスに対して最大限で１５個のブロックＲＡＭが利用可能である。インターフェースバッファに対してブロックＲＡＭリソースを利用可能にするために、これは、ブロックＲＡＭをより小さい数、例えば、１４個のブロックＲＡＭに制限することをもたらす。この例示的で限定ではないリソース割り当ては、パターン断片の長さと、パターン断片および許容可能な誤りの組み合わせとに制約を生じる可能性がある。これは、次に、最大誤りレベルに制限を生じる。「ｍ」がパターンの大きさと等しく、「ｐ」がパターン断片の数に等しく、「ｋ」が許容可能な誤りの数または編集距離である場合、以下の式：

$m_{min} = p_{min}(k+1)$

$m_{max} = p_{max}(k+1)$

【数８】

$$\alpha_{max} = \frac{k}{m_{min}} = \frac{k}{p_{min}(k+1)}$$

が適用可能であり、ここでα＝誤りの数をパターンの大きさで除算した比率である。

【００４０】

以下のテーブル１は、「α」、即ち、誤りの数をパターンの大きさで除算した比率が１以下である場合のものである。

【表１】

テーブル1

k	m_{min}	m_{max}
0	1	14
1	2	28
2	3	42
3	4	56
…	…	…

【００４１】

以下のテーブル２は、「α」、即ち、誤りの数をパターンの大きさで除算した比率が１／２以下である場合のものである。

【表２】

テーブル2		
k	m_{min}	m_{max}
0	2	15
1	4	30
2	6	45
3	8	60
…	…	…

【００４２】

以下のテーブル３は、「α」、即ち、誤りの数をパターンの大きさで除算した比率が１／３以下である場合のものである。

【表３】

テーブル3		
k	m_{min}	m_{max}
0	3	16
1	6	32
2	9	48
3	12	64
…	…	…

【００４３】

それゆえに、第２の例およびテーブル２で、「α」、即ち、誤りの数をパターンの大きさで除算した比率が１／２以下である場合で、パターンが誤りを許容しない、即ち、「ｋ＝０」である場合、少なくとも２個の文字が存在し、１５個以下の文字であるはずである。１個の誤りを許容するパターン、即ち、「ｋ＝１」は、少なくとも４個の文字で、３０個以下の文字を含むはずである。広範な種類の許容パターンの大きさおよび許容可能な誤りが利用され得るが、少なくとも２個の文字が存在して１５個以下の文字であるパターンが、英語での大部分のテキスト検索に対して有効な制約であると考えられており、しかしながら、これが限定として解釈されるべきではない。

【００４４】

断片が許容可能な長さの範囲に一様に分布していると仮定することによって、大まかな能力の推定が展開され得る。例示的で限定ではない例では、各ブルームフィルタ配列５４が約３０００個のパターン断片の能力を有する場合、システムは４２０００個のパターン断片の総合能力を有する。各パターンが３個のパターン断片に分割され得ると仮定される場合、システムは１４０００個のパターンの能力を有する。

【００４５】

図３に示されるように、一度、１つまたは複数のパターン断片長５６に関して可能性のある合致が検出されると、「ｒ」個のパターンのうちの１つに関する近似合致が存在するか否かを判定するために検証ステージ４２によってテキストのその範囲が検査される必要がある。「ｒ」はおよそ１００００パターン以上であり得るので、検証ステージ４２が実行する必要のある検索範囲を削減する必要性が存在する。これこそ、検証ステージ４２が考慮する、可能性のある合致パターン（パターンセット）のセットを低減する重要な役割を削減ステージ４０が提供することである。パラメータが制約の範囲内にある限り、許容可能な誤りの数が各パターンに対して特定され得るという仮定が存在する。

【００４６】

図３に示されるようなこの特許出願全体で、図３に示されるようにフィルタステージ３８、削減ステージ４０および／または検証ステージ４２は、少なくとも１つの再構成可能なロジックデバイス、例えば、フィールドプログラマブルゲートアレイ（「ＦＰＧＡ」）または少なくとも１つの集積回路、例えば、特定用途向け集積回路（「ＡＳＩＣ」）を使

用し得る。

【００４７】

削減ステージ４０を実行する２つの例示的だが限定ではない手法が存在する。第１の手法は、図６の図番１２０で示されるパターンセットを解決するために利用されるデータ検索を単純化することである。これは、データ文字列がシフトレジスタ５２に入ることを可能にし、シフトレジスタ５２は次にブルームフィルタ配列５４を含むことが好ましいフィルタステージまたは回路３８に渡す。その目的はこの手法を用いて、フィルタステージまたは回路３８のブルームフィルタ配列５４によって計算されたハッシュ値のうちの、全てではない場合には一部を、テーブル１２８へのインデックス、例えば、ビンインデックス（ＢｉｎＩｎｄｅｘ）１２４として利用することである。

【００４８】

例えば、パターン断片１２１がフィルタステージまたは回路３８によって受信可能であり、ビンインデックス１２４を含むハッシュ値１２６として受信される。この第１のテーブル１２８へのエントリは、パターン断片に対する識別子１２７、例えば、ビンインデックス１２４を含むハッシュ値１２６をマッピングする断片ＩＤを含む。例えば、例示的な識別子、例えば、例示的なパターン断片１２１と関連付けられている断片$ＩＤ_1$および断片$ＩＤ_4$が存在する。パターン断片に対する識別子１２７、例えば、断片ＩＤは、各パターン断片に割り当てられている固有の２進数タグである。

【００４９】

第２のテーブル１３２は、１つまたは複数のパターン識別子を可能性のある合致パターンのセットに関してインデックス付けするために、パターン断片に対する識別子１２８、例えば、断片ＩＤを利用する。１つまたは複数のパターンが特定のパターン断片を特定し得るので、第２のテーブル１３２のエントリは、１つまたは複数のパターン識別子、例えば、ＰＩＤｓを含む。パターン識別子、例えば、ＰＩＤｓは、各パターンに関連付けられている固有の２進数タグである。例えば、例示的な識別子１３１は、２つのパターン１３７および１３８に関連する。

【００５０】

第３のテーブル１３４は、１つまたは複数の（パターン、許容可能な誤り）対をインデックス付けするためにパターン識別子、例えば、ＰＩＤを利用する。次に、各々の合致パターンが所定の許容可能な誤りを有する、識別されたセットまたは複数の可能性のある合致パターンが、例えば、（パターン、許容可能な誤り）対１３７および１３８が、図番１３６で示されるように生成される。

【００５１】

可能性のある合致１３６のこのパターンセットが検証ステージ４２に渡され、それは合致パターンおよび関連する所定の許容可能な誤りに関する近似合致エンジン１４２による評価を含む。周期当たりの参照数がメモリによってサポートされる参照数を超過しない限りは、そこでは、パターン断片識別子１２８、例えば、断片ＩＤｓ、パターン識別子１３２、例えば、ＰＩＤｓおよびパターン１３４に関するテーブルの１つのコピーだけを必要とする。

【００５２】

削減ステージ４０に関する第２および好適な方法論が存在し、図７の図番１５０で概略的に示されている。これは、データ文字列が、好ましくはブルームフィルタ配列５４を含むフィルタステージまたは回路３８に渡されることを可能にする。この手法を用いると、ブルームフィルタ配列５４での合致を生じる実際のデータセグメントまたは断片が、それらのパターン断片を特定するパターン１６２に対してパターン識別子を解決するために利用される。合致を生じるデータセグメントまたは断片が、図番１５６、１５８および１６０で示されるデータ構造の１つまたは複数のエントリを特定するために使用され、データ構造は、関連するパターン断片を特定するパターンに関するパターン識別子、例えば、ＰＩＤｓを記憶している。適切なデータ構造の例示的であるが、限定ではない例は、ハッシュテーブルおよび平衡探索木である。この方法論は、１つまたは複数のデータ構造を含み

得る。図７の例示的であるが、限定ではない例１５０では、１つのデータ構造が各パターン断片長に対して割り当てられている。

【００５３】

例示的であるが、限定ではない例では、ブルームフィルタ配列５４での合致を生じる２つのデータセグメントまたは断片が図番１２１および１２３で特定される。データ断片１２１は、データ構造のエントリ１５７を削減ステージ４０の一部として特定する。これらのデータ構造１５６、１５８および１６０は、多種多様の異なる構造、例えば、決定木、ハッシュテーブル等などを含み得る。これらの参照（単数または複数）の結果は、パターンセットのパターンに対するパターン識別子のセットである。従来技術の手法と同様に、これらのパターン識別子、例えば、ＰＩＤｓ、１５６、１５８および１６０は、検証ステップの前にテーブル４０からパターンおよび関連する許容可能な誤りを取り出すために利用される。前に参照したエントリ１５７は、可能性のある合致１６４のセットを生じさせるために、パターン１６３および１６５と関連する許容可能な誤りとを特定する。可能性のある合致および関連する所定の許容可能な誤り１６４のセットは、次に、検証ステージ４２の近似合致エンジン１６６によって評価される。

【００５４】

この手法は、パターン識別子のセットを、２つのステップの代わりに単一のステップで解決し、さらに、ブルームフィルタ配列５４によって生じる誤検出誤りを除去する。さらに、パターン識別子構造（単数または複数）のエントリの位置を決定するために実際のデータセグメントが利用されているため、明示的な合致が実行される。テーブルにエントリが存在しない場合、誤検出が検出され、検出ステージ４２にその特定のパターン断片長に対するパターン識別子、例えば、ＰＩＤｓは渡されない。その代償は、データ構造がより複雑である可能性があること、および実装によってはその実装がよりリソース消費型になることである。

【００５５】

それゆえに、これは、ハードウェア実装の影響を受け易い複数のパターンに対する近似パターン合致に関するフィルタ回路３８および削減ステージ４０に関するスケーラブルな設計である。何千ものパターンに加え、複数のフィルタ回路が、周期当たりの複数の入力シンボルをサポートすることが可能である。高性能フィルタ回路３８および削減ステージ４０を利用することによって、検証ステージ４０に配置された性能要求が解析され得る。この解析を目的として、出願人は、全てのパターンが同じ数「ｋ」の許容可能な誤りを特定すると仮定する。検証ステージでの有効負荷は、合致の確率および可能性のある合致パターンのセットの期待される大きさとして決定される。合致の確率は、単純に、「ｒ（ｋ＋１）」個の断片のうちのいずれかがテキスト窓で合致を生じる確率と、ブルームフィルタ配列５４の誤検出確率との総和であり、ここで「ｒ」は「ｒ」個のパターン断片の所定の数であり、「ｋ」は所定の誤りの数である。

【００５６】

「Ｌ」がブルームフィルタ配列５４の数であり、「σ」個の文字のアルファベットに対して無作為のテキストを仮定する場合、これらの断片のいずれかの合致の確率は、

【数９】

$$E[\text{合致}] = \frac{r(k+1)}{\sigma^{\left\lfloor \frac{m}{k+1} \right\rfloor}}$$

になる。

【００５７】

ブルームフィルタ配列の誤検出確率の付加は、

【数１０】

$$E[\text{合致}]=\frac{r(k+1)}{\sigma^{\left\lfloor\frac{m}{k+1}\right\rfloor}}+Lf$$

を提供する。

【００５８】

例示的だが、限定ではない例示値Ｌ＝１４、σ＝４０、ｒ＝１４０００およびｆ＝０．００３４を利用することによって、合致確率は最小断片長に非常に敏感である。例えば、ｍ＝５およびｋ＝１（２個の文字の最小断片長）である場合、合致確率が１である。パターンの大きさが６（３個の文字の最小パターン断片長）に増大されると、合致確率は０．０７９である。この結果は、最小パターン断片の大きさが３個以上の文字であるべきであることを示唆している。この状況では、１２周期毎に１つの合致が期待される。

【００５９】

図６に示される削減ステージ４０の方法論を、一連のインデックス参照と共に利用することによって、もし少なくとも１つのブルームフィルタ配列５４が合致を生じる場合、合致を生じるブルームフィルタ配列５４の期待数、即ち、ビンインデックス参照の期待数は：

【数１１】

$$E[\text{ビン}]\le 1+\frac{L}{\sigma^{\left\lfloor\frac{m}{k+1}\right\rfloor}}+Lf$$

である。

【００６０】

パターン断片が「Ｌ」個のブルームフィルタ配列５４（パターン断片長）上に一様に分布し、ビンの上に一様に分布しているとの仮定の下で、図６に示されるように、ビン当たりのパターン断片識別子１２８の期待数は：

【数１２】

$$E[\text{断片ID/ビン}]\le\frac{r(k+1)}{LW\left(\frac{B}{W}\right)^{b}}$$

である。ここで「Ｂ」は、ブルームフィルタ配列５４を実装するために使用されているメモリの大きさであり、「Ｗ」は、ブルームフィルタ配列５４のワードの数であり、「ｂ」は、ブルームフィルタ配列５４の各ブルームフィルタで使用されているハッシュ関数の数である。

【００６１】

最後に、所与のパターン断片１３１を特定するパターンの期待数は、図６に示されるように：

【数１３】

$$E[\text{パターン/断片ID}]\le 1+\frac{r}{\sigma^{\left\lfloor\frac{m}{k+1}\right\rfloor}}$$

を含む。

【００６２】

それゆえに、少なくとも１つのブルームフィルタ配列５４が合致するという条件で、期待されるパターンの大きさは：

【数１４】

$$E[\text{パターン}] \le \left(1 + \frac{L}{\sigma^{\lfloor m/(k+1) \rfloor}} + Lf\right)\left(\frac{r(k+1)}{LW\left(\frac{B}{W}\right)^{b}}\right)\left(1 + \frac{r}{\sigma^{\left\lfloor \frac{m}{k+1} \right\rfloor}}\right)$$

である。

【００６３】

一様なテキスト、断片長の一様な分布、一様な分布（優良なハッシュ関数）と、Ｌ＝１４、σ＝４０、ｒ＝１４０００、ｂ＝３およびＷ＝５１２とを仮定すると、ｍおよびｋ（パターン長および許容可能な誤り）の実際的な値に対して期待されるパターンセットの大きさは１０未満である。アルファベットの大きさおよび／またはパターンの大きさが増大するに従って、期待されるパターンセットの大きさは、急速に１に近づく。

【００６４】

図７に示される他の削減ステージ手法を、テキストセグメントをインデックスとして用いて使用する場合の期待されるパターンセットの大きさの式は同様であり、若干小さいパターンセットサイズを生じる。１２の周期毎に１つのパターンが合致する前の結果と組み合わせて、検証ステージ４２の平均スループットに対する控えめの制約は、毎周期当たり１つのパターンで約１個の合致である。

【００６５】

従って、新規の発明の複数の実施形態が示され、説明されてきた。前述の説明から明らかであるように、本発明の所定の態様は、本明細書で例示されている例の特定の詳細によって制限されず、それゆえに、当業者が他の修正および適用またはそれらの同等物を創案することが意図されている。前述の明細書で使用されている用語「有する」、「有している」、「含む」および「含んでいる」ならびに類似の用語は、「必要とする」の意味ではなく、「任意選択で」または「含んでよい」の意味として使用されている。一方、本構成の多くの変更、修正、変形および他の使用および用途が、明細書および添付の図面を熟考した後に当業者に明白になるであろう。本発明の精神および範疇から逸脱しないそのような変更、修正、変形および他の使用ならびに用途の全ては、添付の特許請求の範囲によってのみ制限される本発明によって網羅されているとみなされる。

【図面の簡単な説明】

【００６６】

【図１】従来技術の近似パターン合致技術の例示的な概略図である。

【図２】データソース、フィルタ回路、削減ステージおよび検証ステージを含む本発明の例示的なブロック図である。

【図３】合致検出機能、削減ステージならびに近似パターン合致を有する検証ステージを備えるブルームフィルタを含む、本発明の例示的であり、限定ではないブロック図である。

【図４】ブルームフィルタ配列を使用する第１のフィルタリングステージ処理技術の例示的なブロック図である。

【図５】単一のハッシュ関数ジェネレータと共にブルームフィルタを使用する第２のフィルタリングステージ処理技術の例示的なブロック図である。

【図６】第１の削減ステージ処理技術の例示的なブロック図である。

【図７】第２の削減ステージ処理技術の例示的なブロック図である。

【図１】

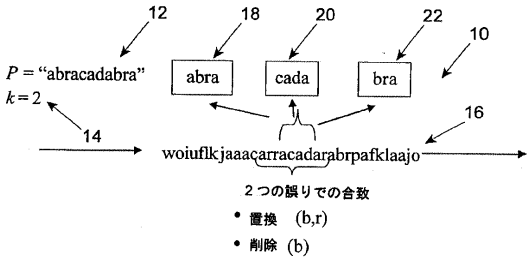

FIG. 1

【図２】

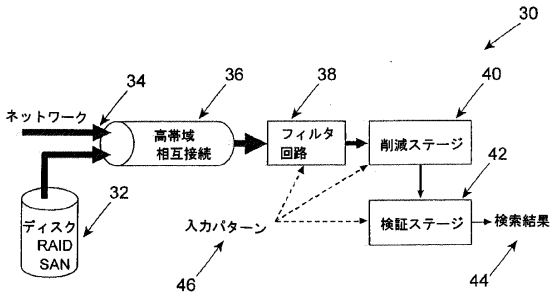

FIG. 2

【図３】

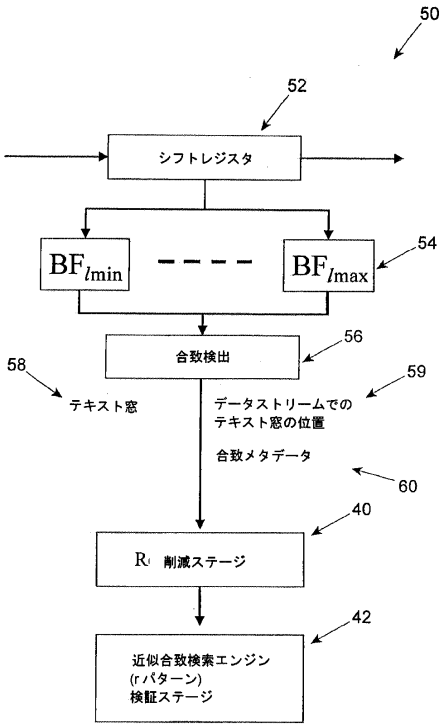

FIG. 3

【図４】

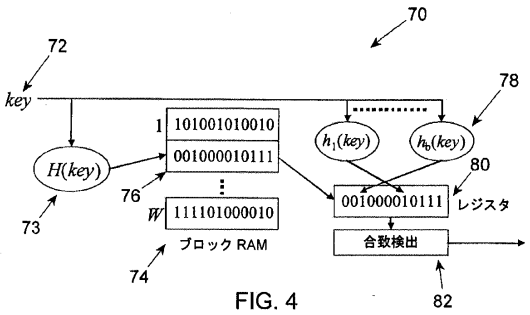

FIG. 4

【図５】

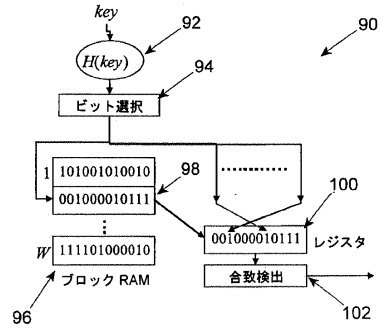

FIG. 5

【図６】

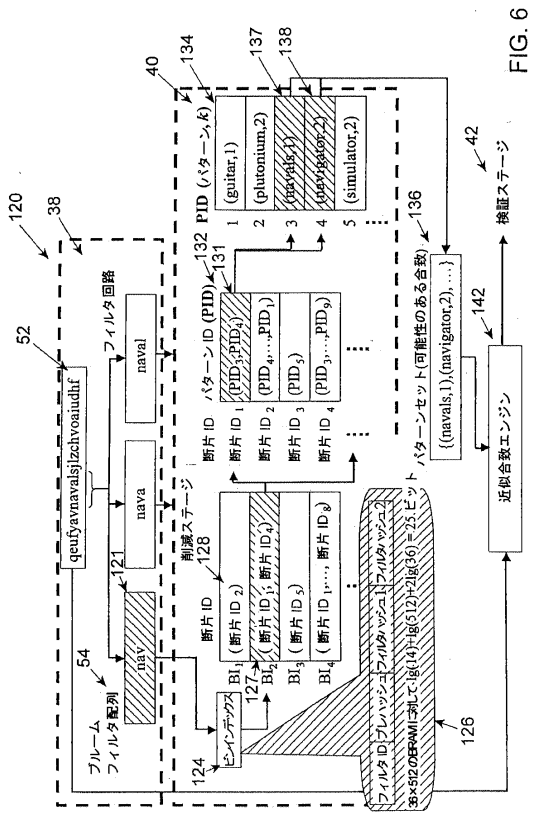

FIG. 6

【図７】

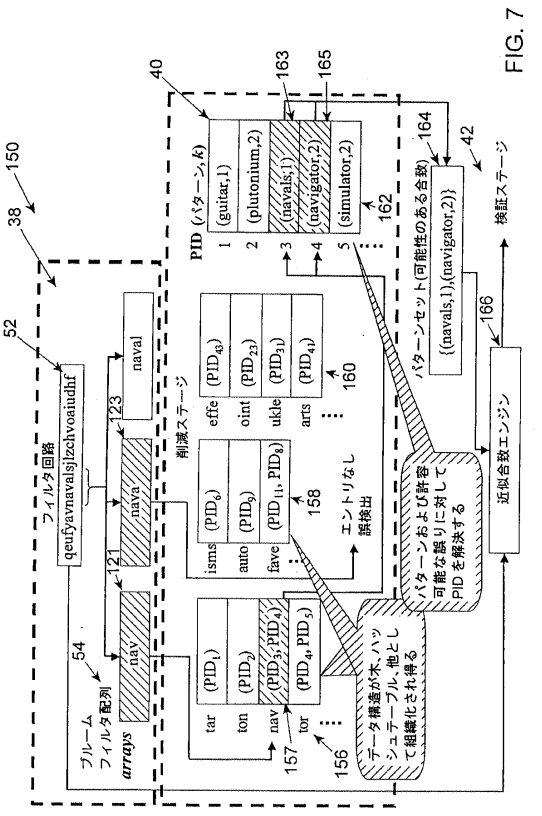

FIG. 7

フロントページの続き

(74)代理人 100062007
弁理士 川口 義雄
(72)発明者 テイラー，デイビツド・エドワード
アメリカ合衆国、ミズーリ・６３１１８、セント・ルイス、ミズーリ・アベニユー・３４４８

審査官 早川 学

(56)参考文献 特表２００８－５３２１７７（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０