(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2013-196665
(P2013-196665A)
(43)公開日　平成25年9月30日(2013.9.30)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	350C	
		G06F	17/30	415	

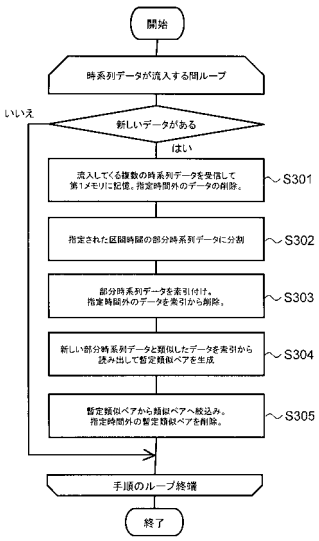
審査請求 未請求　請求項の数 5　ＯＬ　（全 10 頁）

(21)出願番号　特願2012-66643(P2012-66643)
(22)出願日　平成24年3月23日(2012.3.23)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(74)代理人　100119677
弁理士　岡田　賢治
(74)代理人　100115794
弁理士　今下　勝博
(72)発明者　倉沢　央
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　佐藤　浩史
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

(54)【発明の名称】データ検索装置、データ検索方法、及びデータ検索プログラム

(57)【要約】
【課題】本発明の目的は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、類似した部分時系列データのペアを検索するに際し、メモリサイズを抑えることができるデータ検索装置、データ検索方法、及びデータ検索プログラムを提供することである。
【解決手段】本発明では、逐次入力される複数系列の時系列データを所定時間の部分時系列データに分割し、前記部分時系列データについて互いの類似度を求め、前記類似度に基づいて、指定された時間範囲の受信時刻毎に、類似した部分時系列データのペアの候補と類似度の閾値を記憶しておき、生成された部分時系列データのペアの中から、記憶した類似度の閾値よりも類似した部分時系列データを暫定類似ペアとして記憶しておき、記憶した暫定類似ペアの中から、与えられた数の最も類似した部分類似時系列データペアを読み出す。
【選択図】図３



【特許請求の範囲】
【請求項１】
逐次流入してくる複数系列の時系列データを受信する受信部と、
前記受信部が受信した前記時系列データを記憶する第１メモリと、
前記第１メモリから時系列データを読み出し、所定時間の部分時系列データに分割する区間分割部と、
前記区間分割部が分割した前記部分時系列データについて互いの類似度を求める類似判定部と、
前記区間分割部で生成される前記部分時系列データのペアの中から前記類似判定部が求めた類似度に基づき、所定の類似度の閾値よりも類似した部分時系列データを暫定類似ペア候補として検索して記憶させる暫定類似ペア候補検索部と、
前記暫定類似ペア候補検索部の検索する前記暫定類似ペア候補及び前記所定の類似度の閾値を記憶する第３メモリと、
前記第３メモリから読み出された前記暫定類似ペア候補から、類似度の閾値判定と類似ペアの絞込みをする類似ペア判定部と
を備えるデータ検索装置。
【請求項２】
前記暫定類似ペア候補検索部が、
前記区間分割部で生成される前記部分時系列データのペアの中から前記類似判定部が求めた類似度に基づき、部分時系列データを受信時刻で区別可能な構造を有した類似検索索引を作成する索引部と、
前記索引部で作成された類似検索索引を記憶する第２メモリと、
を備えることを特徴とする請求項１に記載のデータ検索装置。
【請求項３】
逐次入力される複数系列の時系列データを所定時間の部分時系列データに分割し、前記部分時系列データについて互いの類似度を求め、前記類似度に基づいて、指定された時間範囲の受信時刻毎に、類似した部分時系列データのペアの候補と類似度の閾値を記憶しておき、生成された部分時系列データのペアの中から、記憶した類似度の閾値よりも類似した部分時系列データを暫定類似ペアとして記憶しておき、記憶した暫定類似ペアの中から、与えられた数の最も類似した部分類似時系列データペアを読み出すデータ検索方法。
【請求項４】
前記部分時系列データについて互いの類似度を求める際、前記部分時系列データの受信時刻で区別可能な構造を有した類似検索索引を使って、前記部分時系列データの受信時刻毎の類似度の閾値で絞り込むことを特徴とする請求項３に記載のデータ検索方法。
【請求項５】
請求項３又は４に記載のデータ検索方法をコンピュータに実行させるデータ検索プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、類似した部分時系列データのペアを検索するデータ検索装置、データ検索方法、及びデータ検索プログラムに関する。
【背景技術】
【０００２】
時系列データは、時間的に連続して観測されたデータ系列である。時系列データは環境観測や設備監視、株価、統計情報、メディア処理などで利用されている。
【０００３】
時系列データの検索には、事前に蓄積されたデータに対して検索するものと、逐次流入してくるデータに対して検索するものとがある。前者の例としては、過去に定時観測した

降水量のデータで類似したパターンを探す処理などがある。後者の例としては、リアルタイムに株価の変動を観測して目的のパターンと一致した瞬間に通知する処理などがある。本発明は後者のリアルタイムな処理を扱う。

【０００４】

リアルタイムな時系列データ検索の先行技術としては、以下のものが挙げられる。複数系列の同一時刻の時系列データのうち、類似した時系列データのペアを求める検索処理があった（例えば、特許文献１や非特許文献１参照）。特許文献１では閾値よりも類似度の高いペアを、非特許文献１では相関係数が閾値以上のものを探す課題を扱っている。これらの先行技術では、系列数が多いほど類似判定の計算コストが大きくなるため、類似度計算の削減のための次元圧縮や類似度計算回数削減のための索引技術を技術課題としていた。しかしながら、これらの先行技術は時系列データの終点が現在時刻のもののみを扱うという制限をもつ。つまり、時間的なずれのある類似した部分時系列データを対象とした検索には適用できない。

【０００５】

一方、現在から過去の一定期間内に流入した単一系列の時系列データの中から、類似した部分時系列データのペアを求める検索処理があった（例えば、非特許文献２と非特許文献３参照）。非特許文献２と非特許文献３のいずれも、指定した数の最も類似した部分時系列データのペアを検索する課題を扱っている。これらの先行技術では、検索対象とする時間が長い、もしくは、時系列データの長さに対して部分時系列データが短いほど暫定類似ペアを記憶するメモリサイズが大きくなるため、暫定類似ペア候補の絞り込み方法を技術課題としていた。しかしながら、これらの先行技術は単一系列の時系列データのみを扱うという制限を持つ。つまり、複数系列の時系列データを対象とした検索には適用できない。

【先行技術文献】

【特許文献】

【０００６】

【特許文献１】特開２００６－２４４３８９号公報

【非特許文献】

【０００７】

【非特許文献１】Ａｂｄｕｌｌａｈ Ｍｕｅｅｎ, Ｓｕｍａｎ Ｎａｔｈ ａｎｄ Ｊｉｅ Ｌｉｕ, “Ｆａｓｔ Ａｐｐｒｏｘｉｍａｔｅ Ｃｏｒｒｅｌａｔｉｏｎ ｆｏｒ Ｍａｓｓｉｖｅ Ｔｉｍｅ－ｓｅｒｉｅｓ Ｄａｔａ”, Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ２０１０ ＡＣＭ ＳＩＧＭＯＤ ｉｎｔｅｒｎａｔｉｏｎａｌ ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｍａｎａｇｅｍｅｎｔ ｏｆ ｄａｔａ, ２０１０.

【非特許文献２】Ａｂｄｕｌｌａｈ Ｍｕｅｅｎ ａｎｄ Ｅａｍｏｎｎ Ｋｅｏｇｈ, “Ｏｎｌｉｎｅ Ｄｉｓｃｏｖｅｒｙ ａｎｄ Ｍａｉｎｔｅｎａｎｃｅ ｏｆ Ｔｉｍｅ Ｓｅｒｉｅｓ Ｍｏｔｉｆｓ”, Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ １６ｔｈ ＡＣＭ ＳＩＧＫＤＤ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｋｎｏｗｌｅｄｇｅ Ｄｉｓｃｏｖｅｒｙ ａｎｄ Ｄａｔａ Ｍｉｎｉｎｇ, ２０１０.

【非特許文献３】Ｈｏａｎｇ Ｔｈａｎｈ Ｌａｍ, Ｔｕ Ｅｉｎｄｈｏｖｅｎ, Ｔｏｏｎ Ｃａｌｄｅｒｓ ａｎｄ Ｎｉｎｈ Ｐｈａｍ, “Ｏｎｌｉｎｅ Ｄｉｓｃｏｖｅｒｙ ｏｆ Ｔｏｐ－ｋ Ｓｉｍｉｌａｒ Ｍｏｔｉｆｓ ｉｎ Ｔｉｍｅ Ｓｅｒｉｅｓ Ｄａｔａ”, Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ２０１１ ＳＩＡＭ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｄａｔａ Ｍｉｎｉｎｇ, ２０１１.

【非特許文献４】Ｙｉａｎｉｌｏｓ, Ｐｅｔｅｒ Ｎ, “Ｄａｔａ ｓｔｒｕｃｔｕｒｅｓ ａｎｄ ａｌｇｏｒｉｔｈｍｓ ｆｏｒ ｎｅａｒｅｓｔ ｎｅｉｇｈｂｏｒ ｓｅａｒｃｈ ｉｎ ｇｅｎｅｒａｌ ｍｅｔｒｉｃ ｓｐａｃｅｓ”, Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ｆｏｕｒｔｈ ａｎｎｕａｌ ＡＣＭ－ＳＩＡＭ Ｓｙｍｐｏｓｉｕｍ ｏｎ Ｄｉｓｃｒｅｔｅ ａｌｇｏｒｉｔｈｍｓ. Ｓｏｃｉｅｔｙ

for Industrial and Applied Mathematics Philadelphia, pp. 311−321, 1993.

【発明の概要】

【発明が解決しようとする課題】

【０００８】

上述のように、従来技術は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、類似した部分時系列データのペアを求める検索処理を扱うことができない。つまり、時間的なずれのある類似した部分時系列データを複数系列の時系列データの中から検索することができない。このような検索例としては、１ヶ月間の複数銘柄の株価の時系列データから類似した部分時系列データを検索して、時間的なずれのある特徴的な株価の変動を見つけ出す処理がある。また、複数のセンサ値の時系列データから類似した部分時系列データを検索して、日射から地熱、気温といった時間的なずれの生じる現象を見つけ出す処理も検索例として挙げられる。上述のとおり、複数系列の時系列データを対象とした先行技術や、時間的なずれのある類似した部分時系列データを対象とした先行技術はあるが、この２つを同時に満たす時系列データ検索処理技術は見当たらない。

【０００９】

上記２つを同時に満たす時系列データ検索処理では、現在から過去の一定期間内に流入した複数系列の時系列データの中から、与えられた数の最も類似した部分時系列データのペアを探す検索処理を扱うことになる。この検索処理は、メモリサイズを抑えることが求められており、暫定類似ペア候補の絞り込みが解決すべき技術課題となっている。ここで、「与えられた数の最も類似した部分時系列データのペアを探す検索処理」とは、「要求されている出力数ｋ（ペアの個数）が入力されたとき、最類似からｋ番目までの部分時系列データのペアを検索して出力する処理」を意味する。

【００１０】

与えられた数の最も類似した部分時系列データのペアを探す検索要求を扱うとき、検索対象とする時系列データの時間が長い、もしくは時系列データの長さに対して部分時系列データが短いほど、暫定類似ペアの数が増える。暫定類似ペアの数が多いと、それらを記憶するメモリサイズが大きくなる。また、暫定類似ペアを絞りこめないと、暫定類似ペア間の類似度の閾値を大きく見積もってしまうことになり、検索索引を使った枝刈りが難しく、距離計算回数が大きくなる。以上の理由から、メモリサイズを抑えるためには暫定類似ペアの絞り込みが必要という課題がある。ここで、「枝刈り」とは、検索対象とならないデータを予め除去しておくという意味である。

【００１１】

本発明は、上記課題を解消するためになされたものであり、その目的は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、類似した部分時系列データのペアを検索するに際し、必要とするメモリサイズを抑えることができるデータ検索装置、データ検索方法、及びデータ検索プログラムを提供することにある。

【課題を解決するための手段】

【００１２】

本発明は、新しい受信時刻の部分時系列データの挿入と古い受信時刻の部分時系列データの削除が容易なデータ構造で、かつ、部分時系列データを受信時刻ごとに区別して枝刈りするアルゴリズム採用することとした。

【００１３】

本発明に係るデータ検索装置は、逐次流入してくる複数系列の時系列データを受信する受信部と、前記受信部が受信した前記時系列データを記憶する第１メモリと、前記第１メモリから時系列データを読み出し、所定時間の部分時系列データに分割する区間分割部と、前記区間分割部が分割した前記部分時系列データについて互いの類似度を求める類似判定部と、前記区間分割部で生成される前記部分時系列データのペアの中から前記類似判定部が求めた類似度に基づき、所定の類似度の閾値よりも類似した部分時系列データを暫定類似ペア候補として検索して記憶させる暫定類似ペア候補検索部と、前記暫定類似ペア候

補検索部の検索する前記暫定類似ペア候補及び前記所定の類似度の閾値を記憶する第３メモリと、前記第３メモリから読み出された前記暫定類似ペア候補から、類似度の閾値判定と類似ペアの絞込みをする類似ペア判定部と、を備える。

【００１４】

本発明に係るデータ検索方法は、逐次入力される複数系列の時系列データを所定時間の部分時系列データに分割し、前記部分時系列データについて互いの類似度を求め、前記類似度に基づいて、指定された時間範囲の受信時刻毎に、類似した部分時系列データのペアの候補と類似度の閾値を記憶しておき、生成された部分時系列データのペアの中から、記憶した類似度の閾値よりも類似した部分時系列データを暫定類似ペアとして記憶しておき、記憶した暫定類似ペアの中から、与えられた数の最も類似した部分類似時系列データペアを読み出す。

【００１５】

本発明は、指定された時間範囲の受信時刻毎に暫定類似ペアを管理することで、古い受信時刻で暫定類似ペア候補には含まれるが類似ペアの閾値を満たさない部分時系列データのペアの判定と絞込みを高速にできるようになる。このため、本発明は、検索対象とする時系列データの時間が長い、もしくは時系列データの長さに対して部分時系列データが短いときに、メモリサイズの削減を実現することができる。

【００１６】

従って、本発明は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、類似した部分時系列データのペアを検索するに際し、メモリサイズを抑えることができるデータ検索装置及びデータ検索方法を提供することができる。

【００１７】

本発明に係るデータ検索装置は、前記暫定類似ペア候補検索部が、前記区間分割部で生成される前記部分時系列データのペアの中から前記類似判定部が求めた類似度に基づき、部分時系列データを受信時刻で区別可能な構造を有した類似検索索引を作成する索引部と、前記索引部で作成された類似検索索引を記憶する第２メモリと、を備える。

【００１８】

本発明に係るデータ検索方法は、前記部分時系列データについて互いの類似度を求める際、前記部分時系列データの受信時刻で区別可能な構造を有した類似検索索引を使って、前記部分時系列データの受信時刻毎の類似度の閾値で絞り込む。

【００１９】

従って、本発明は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、類似した部分時系列データのペアを検索するに際し、必要とする距離計算コストを抑えることができるデータ検索装置及びデータ検索方法を提供することができる。

【００２０】

本発明に係るデータ検索プログラムは、前記データ検索方法をコンピュータに実行させる。

【発明の効果】

【００２１】

本発明は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、類似した部分時系列データのペアを検索するに際し、メモリサイズを抑えることができるデータ検索装置及びデータ検索方法を提供することができる。

【図面の簡単な説明】

【００２２】

【図１】類似部分時系列ペアを説明するイメージの図である。

【図２】本発明に係るデータ検索装置を説明する図である。

【図３】本発明に係るデータ検索方法を説明する図である。

【図４】本発明に係るデータ検索装置が時系列データを部分時系列データに分割するときのイメージの図である。

【図５】類似ペアの絞込みのイメージ図である。

【発明を実施するための形態】

【００２３】

添付の図面を参照して本発明の実施形態を説明する。以下に説明する実施形態は本発明の実施例であり、本発明は、以下の実施形態に制限されるものではない。なお、本明細書及び図面において符号が同じ構成要素は、相互に同一のものを示すものとする。

【００２４】

図１は、類似部分時系列ペアのイメージの図である。実施形態の装置では与えられた数の最も類似した部分類似時系列データペアを検索する。

【００２５】

図２は、本実施形態のデータ検索装置３０３を説明する図である。データ検索装置３０３は、逐次流入してくる複数系列の時系列データを受信する受信部１１と、受信部１１が受信した前記時系列データを記憶する第１メモリ１２と、第１メモリ１２から時系列データを読み出し、所定時間の部分時系列データに分割する区間分割部１３と、区間分割部１３が分割した前記部分時系列データについて互いの類似度を求める類似判定部１４と、区間分割部１３で生成される前記部分時系列データのペアの中から類似判定部１４が求めた類似度に基づき、所定の類似度の閾値よりも類似した部分時系列データを暫定類似ペア候補として検索して記憶させる暫定類似ペア候補検索部１９と、暫定類似ペア候補検索部１９の検索する前記暫定類似ペア候補及び前記所定の類似度の閾値を記憶する第３メモリ１８と、第３メモリ１８から読み出された前記暫定類似ペア候補から、類似度の閾値判定と類似ペアの絞込みをする類似ペア判定部１７と、を備える。

【００２６】

図３は、データ検索装置３０３が行うデータ検索方法を説明する図である。本データ検索方法は、逐次入力される複数系列の時系列データを所定時間の部分時系列データに分割し（ステップＳ３０１、Ｓ３０２）、前記部分時系列データについて互いの類似度を求め、前記類似度に基づいて、指定された時間範囲の受信時刻毎に、類似した部分時系列データのペアの候補と類似度の閾値を記憶しておき（ステップＳ３０３）、生成された部分時系列データのペアの中から、記憶した類似度の閾値よりも類似した部分時系列データを暫定類似ペアとして記憶しておき（ステップＳ３０４）、記憶した暫定類似ペアの中から、与えられた数の最も類似した部分類似時系列データペアを読み出す（ステップＳ３０５）。

【００２７】

ｋ個の最も類似した部分時系列データペアを探索する場合の実施例について述べる。

ステップＳ３０１において、受信部１１は、逐次流入してくる複数系列の時系列データを受信して第１メモリ１２に記憶させる。第１メモリ１２の容量に制限がある場合は、指定された時間範囲外の受信時刻の時系列データを第１メモリ１２から削除する。第１メモリ１２の最低限必要な記憶容量は、検索対象とする部分時系列データ長をｍ、検索対象とする時間長をｗ、系列数をｎとすると、ｎ（ｍ＋ｗ－１）の要素数を蓄積できる容量となる。

【００２８】

ステップＳ３０２において、区間分割部１３は、第１メモリ１２から時系列データを読みだして、指定された区間時間の部分時系列データに分割する。本実施形態における部分時系列データとは、逐次流入してくる時間的に連続したデータ系列のうち、指定された区間時間の連続したデータ系列をいう。区間分割部１３では新しい時系列データを受信するたびに、各時系列データの直近の長さｍの部分時系列データを分割して、暫定類似ペア候補検索部１９に部分時系列データを入力する。また、類似ペア判定部１７から指定された時間の長さｍの部分時系列データを分割して返す。図４は、区間分割部１３が時系列データを部分時系列データに分割するときのイメージの図である。図中の

【数１】

$$S_{t_x}^{i} \qquad (1)$$

は受信時刻t_xにおける系列のＩＤがｉの部分時系列データを表している。

【００２９】

ステップＳ３０３において、類似判定部１４は部分時系列データ間の類似度を計算する。本実施形態における類似度とは、２つの部分時系列データがどの程度類似しているかを数値で表す指標である。特に、本実施形態においては、２つの部分時系列データ

【数２】

$$S^i_{t_x} \text{ と } S^j_{t_y} \qquad (2)$$

の類似度Ｄｉｓｔにユークリッド距離を採用し、以下の通り定義する。

【数３】

$$\mathrm{Dist}\left(S^i_{t_x}, S^j_{t_y}\right) = \sqrt{\sum_{c=1}^{m}\left(S^i\left[t_x - c\right] - S^j\left[t_y - c\right]\right)^2} \qquad (3)$$

ただし、S^a_b は系列のＩＤがａ、受信時刻がｂ、長さがｍの部分時系列データ、

$S^a[c]$ は系列のＩＤがａ、受信時刻ｃの要素を表す。

【００３０】

なお、本発明のいう類似度は、上記特徴を満たす様々な距離がその対象となるものであって、ユークリッド距離に限られない。具体的に一例を挙げると、Ｄｙｎａｍｉｃ ｔｉｍｅ ｗａｒｐｉｎｇやＥａｒｔｈ Ｍｏｖｅｒ’ｓ Ｄｉｓｔａｎｃｅ、編集距離、ハミング距離、Ｊａｃｃａｒｄ係数距離、コサイン類似度、マンハッタン距離、マハラノビス距離などを含む時系列データ間の類似度を表す距離であってよい。さらに、これらの距離に加えて、部分時系列のオフセットや振幅を調整したものに対する距離であってもよい。

【００３１】

第３メモリ１８は、受信時刻ごとの類似した部分時系列データのペアの候補と類似度の閾値を記憶する。

【００３２】

受信時刻ごとの類似した部分時系列データのペアの候補は暫定類似ペア候補検索部１９によって書き換えられ、受信時刻ごとの類似度の閾値は暫定類似ペア候補検索部１９と類似ペア判定部１７によって書き換えられる。

【００３３】

ステップＳ３０３において、暫定類似ペア候補検索部１９は、与えられた数の最も類似した部分時系列データペアに含まれうる暫定類似ペアと、受信時刻別の類似度の閾値を第３メモリで管理する。暫定類似ペア候補検索部１９は、暫定類似ペアを受信時刻別にヒープ構造で蓄積し、さらにそのヒープを受信時刻の順にキュー構造で管理する。各ヒープは最大ｋ個の部分時系列データのペアとデータ間の類似度を蓄積する。暫定類似ペアに含まれる部分時系列データのペアは、２つの部分時系列データのうち古いほうの受信時刻を基準に整理する。例えば、数２の部分時系列データ（$t_x < t_y$）でのペアを管理する場合、t_xの受信時刻にもとづいて整理する。こうすることで、最も古い受信時刻の暫定類似ペアの削除を容易にし、さらに各受信時刻で最も距離が大きい類似ペアを容易に削除できるようにしている。

【００３４】

ステップＳ３０４において、暫定類似ペア候補検索部１９は、区間分割部１３で分割した最新の受信時刻の部分時系列データをクエリとして、区間分割部１３で読み込める部分時系列データとの距離が受信時刻別の類似度の閾値よりも類似した部分時系列データを探し、暫定類似ペアのヒープのキューに追加する。

【００３５】

データ検索装置３０３では、この類似した部分時系列データを探すコストを削減するため、暫定類似ペア候補検索部１９には索引部１５と第２メモリ１６を採用して類似検索索引を構築してもよい。このとき、索引部１５と第２メモリ１６は、部分時系列データについて互いの類似度を求める際、部分時系列データの受信時刻で区別可能な構造を有した類似検索索引を使って、部分時系列データの受信時刻毎の類似度の閾値で絞り込む。具体的には、索引部１５と第２メモリ１６は、区間分割部１３で分割した最新の受信時刻の部分時系列データを基準とした類似検索索引を１つ新たにつくり、第２メモリ１６の索引を蓄積するキューに追加する。同時に、指定された時間範囲外の受信時刻の類似検索索引を第２メモリ１６の索引を蓄積するキューから削除する。

【００３６】

［類似検索索引の作成方法］

本実施形態における類似検索索引は、距離空間の類似検索索引Ｖａｎｔａｇｅ　Ｐｏｉｎｔ　Ｔｒｅｅ（ＶＰＴ）（非特許文献４参照）を採用し、以下の手順で作成する。この索引は、基準となる部分時系列データ（基準部分時系列データ）１つと分割距離１つを定めて、基準部分時系列データからの距離と分割距離に基づいて索引対象の複数の部分時系列データを２つに分類することを繰り返して作成する。ここで、受信時刻ｔの部分時系列データ集合S_tを２分類する例を説明する。基準となる系列ｐの部分時系列データS^p_tをS_tから一つ選び、S_tを均等２分割できる距離をr^p_tとしたとき、S_tは、

【数４】

$$S_{in} = \left\{ S_t^i \text{ in } S_t \,|\, \mathrm{Dist}\left(S_t^i, S_t^p\right) \le r_t^p \right\}$$
$$S_{out} = \left\{ S_t^i \text{ in } S_t \,|\, \mathrm{Dist}\left(S_t^i, S_t^p\right) > r_t^p \right\} \qquad (4)$$

の２つの部分時系列データ集合S_{in}とS_{out}に分類される。S_{in}とS_{out}も同様にそれぞれで基準となる部分時系列データと分割距離を決めてさらに空間を２分類することを繰り返すことで、木構造の類似検索索引をつくる。この類似検索索引は、葉ノードと内部ノードという２種類のノードで構成される。葉ノードは木構造の末端のノードであり、内部ノードは末端以外に位置するノードをいう。分類のたびに必要となるS^p_tとr^p_tは、内部ノードを生成して管理する。十分に小さい数の部分時系列データ集合まで繰り返し分類して必要な数の内部ノードを生成した後は、葉ノードを生成してS_tを管理する。つまり、内部ノードは分類に必要な情報（S^p_t、r^p_t）を管理し、葉ノードでは部分時系列データ集合を管理する。

【００３７】

本発明の類似検索索引の作成方法はこれに限るものでなく、部分時系列データを受信時刻で区別可能な構造であればよく、複数の類似検索索引をキュー構造で管理せずに１つの類似検索索引に集約することも可能である。

【００３８】

また、本発明の範囲はこれに限るものでなく、類似検索索引を持たないで暫定類似ペア候補を探すことも可能である。さらに、本発明の範囲はこれに限るものでなく、次元圧縮技術で類似した部分時系列データを探すコストを削減することもできる。具体的な一例を挙げると、Ｄｉｓｃｒｅｔｅ　Ｆｏｕｒｉｅｒ　Ｔｒａｎｓｆｏｒｍａｔｉｏｎ（ＤＦＴ）、Ｓｉｎｇｕｌａｒ　Ｖａｌｕｅ　Ｄｅｃｏｍｐｏｓｉｔｉｏｎ（ＳＶＤ）、Ｐｉｅｃｅｗｉｓｅ　Ａｇｇｒｅｇａｔｅ　Ａｐｐｒｏｘｉｍａｔｉｏｎ（ＰＡＡ）、Ｓｙｍｂｏｌｉｃ　Ａｇｇｒｅｇａｔｅ　Ａｐｐｒｏｘｉｍａｔｉｏｎ（ＳＡＸ）などを含む。

【００３９】

受信時刻別の類似度の閾値は、受信時刻別のヒープに蓄積される暫定類似ペアが上限数ｋに達したとき、ヒープに蓄積されている暫定類似ペアの距離の最大値に更新される。

【００４０】

ステップ３０５において、類似ペア判定部１７は、第３メモリ１８から読み出された類

似部分時系列データの暫定類似ペア候補から、類似度の閾値判定と類似ペア候補の絞り込みをする。類似ペア判定部１７は、最大ｋペアの類似ペア候補を管理できるヒープを１つ生成する。そして、新しい受信時刻の暫定類似ペア候補から順に、ヒープに追加する。各受信時刻でヒープを追加した後、受信時刻別の類似度の閾値をヒープの管理するペア間の距離の最大値に設定する。この新しい閾値よりも類似していない暫定類似ペアを受信時刻別暫定類似ペアのヒープから削除する。この手順によって、新しい受信時刻ほど暫定類似ペアの距離の閾値は大きい値が設定される。また、この手順によって、受信時刻別のヒープに蓄積する暫定類似ペアを上限数ｋよりも下回る数に絞り込み、ヒープを記憶する第３メモリに必要なメモリサイズを削減する。

【００４１】

図５は、類似ペアの絞り込みのイメージ図を示す。ｔ－４からｔの範囲の受信時刻で上位３件の類似部分時系列ペアを探す様子を表している。図上段の網模様の部分は受信時刻ｔの部分時系列データを含む、新たに暫定類似ペア候補に追加されたペアを表している。図中段では、類似ペア候補を格納するヒープに、受信時刻の新しい暫定類似ペアを追加していき、ヒープを更新していく様子を表している。図下段は、類似ペア候補のヒープを更新していく過程で生成した暫定類似ペアの距離の閾値を表している。

【００４２】

以上の手順で、データ検索装置３０３は、現在から過去の一定期間内に流入した複数系列の時系列データの中から、与えられた数の最も類似した部分時系列データペアを検索する。

【００４３】

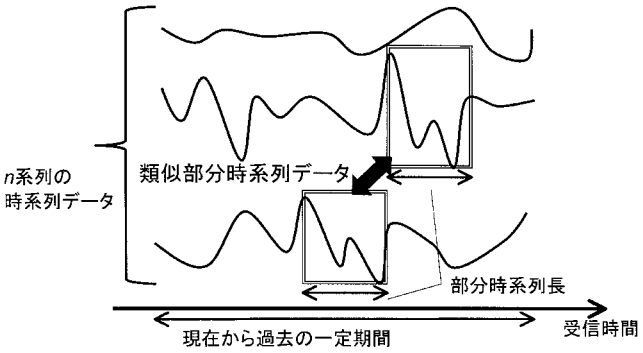
本発明のデータ検索方法をコンピュータに実行させるデータ検索プログラムは、前述した手順をコンピュータに実行させるものである。

【符号の説明】

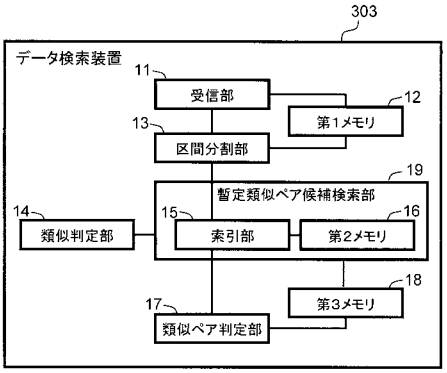
【００４４】

１１：受信部
１２：第１メモリ
１３：区間分割部
１４：類似判定部
１５：索引部
１６：第２メモリ
１７：類似ペア判定部
１８：第３メモリ
１９：暫定類似ペア候補検索部
３０３：データ検索装置

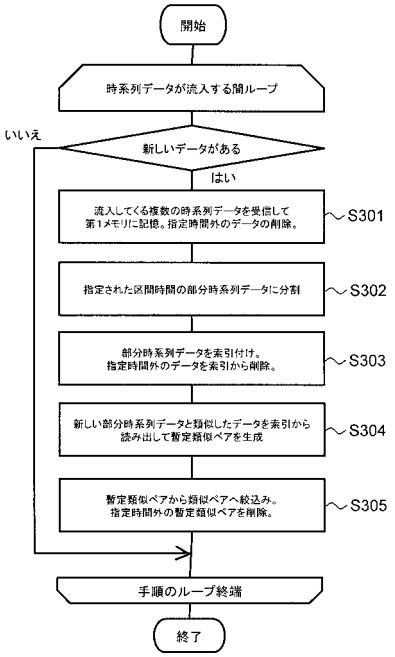
【図１】


n系列の
時系列データ
類似部分時系列データ
部分時系列長
現在から過去の一定期間
受信時間

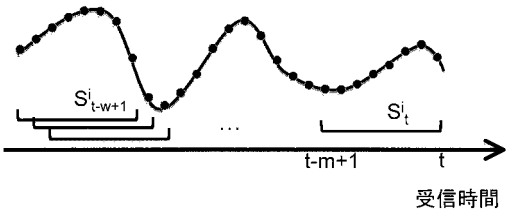
【図２】


303
データ検索装置
11
受信部
12
第1メモリ
13
区間分割部
19
暫定類似ペア候補検索部
14
類似判定部
15
索引部
16
第2メモリ
18
第3メモリ
17
類似ペア判定部

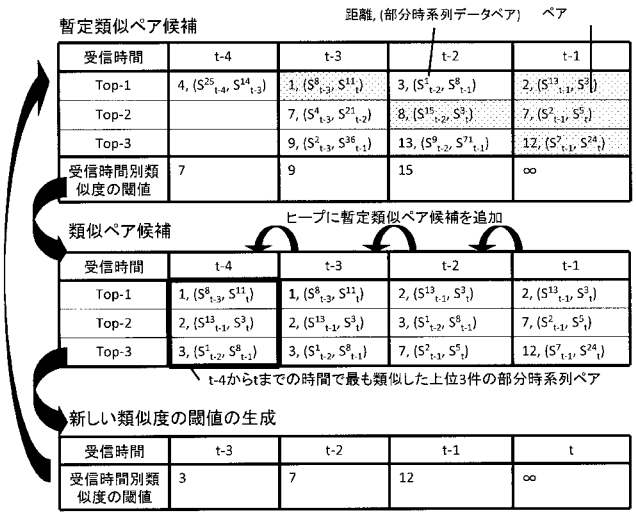
【図３】


開始
時系列データが流入する間ループ
いいえ
新しいデータがある
はい
流入してくる複数の時系列データを受信して
第1メモリに記憶。指定時間外のデータの削除。
S301
指定された区間時間の部分時系列データに分割
S302
部分時系列データを索引付け。
指定時間外のデータを索引から削除。
S303
新しい部分時系列データと類似したデータを索引から
読み出して暫定類似ペアを生成
S304
暫定類似ペアから類似ペアへ絞込み。
指定時間外の暫定類似ペアを削除。
S305
手順のループ終端
終了

【図４】


S^i_{t-w+1}
…
S^i_t
t-m+1
t
受信時間

【図５】


暫定類似ペア候補
距離, (部分時系列データペア)
ペア
受信時間 | t-4 | t-3 | t-2 | t-1
Top-1 | 4, (S^{25}_{t-4}, S^{14}_{t-3}) | 1, (S^{8}_{t-3}, S^{11}_{t}) | 3, (S^{1}_{t-2}, S^{8}_{t-1}) | 2, (S^{13}_{t-1}, S^{3}_{t})
Top-2 | | 7, (S^{4}_{t-3}, S^{21}_{t-2}) | 8, (S^{15}_{t-2}, S^{3}_{t}) | 7, (S^{2}_{t-1}, S^{5}_{t})
Top-3 | | 9, (S^{2}_{t-3}, S^{36}_{t-1}) | 13, (S^{9}_{t-2}, S^{71}_{t-1}) | 12, (S^{7}_{t-1}, S^{24}_{t})
受信時間別類似度の閾値 | 7 | 9 | 15 | ∞
類似ペア候補
ヒープに暫定類似ペア候補を追加
受信時間 | t-4 | t-3 | t-2 | t-1
Top-1 | 1, (S^{8}_{t-3}, S^{11}_{t}) | 1, (S^{8}_{t-3}, S^{11}_{t}) | 2, (S^{13}_{t-1}, S^{3}_{t}) | 2, (S^{13}_{t-1}, S^{3}_{t})
Top-2 | 2, (S^{13}_{t-1}, S^{3}_{t}) | 2, (S^{13}_{t-1}, S^{3}_{t}) | 3, (S^{1}_{t-2}, S^{8}_{t-1}) | 7, (S^{2}_{t-1}, S^{5}_{t})
Top-3 | 3, (S^{1}_{t-2}, S^{8}_{t-1}) | 3, (S^{1}_{t-2}, S^{8}_{t-1}) | 7, (S^{2}_{t-1}, S^{5}_{t}) | 12, (S^{7}_{t-1}, S^{24}_{t})
t-4からtまでの時間で最も類似した上位3件の部分時系列ペア
新しい類似度の閾値の生成
受信時間 | t-3 | t-2 | t-1 | t
受信時間別類似度の閾値 | 3 | 7 | 12 | ∞