Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

揺動型時系列データに対する高速類似部分検索

山内祥裕[†1,†2] 宝珍輝尚[†1] 野宮浩揮[†1]
中西秀哉[†3] 小嶋護[†3]

本論文では，動きの激しい揺動型の時系列データを対象に，波形の一部分をキーとして高速に類似検索を行う手法を提案する．提案手法では，波形を全体から見ると極小の長さの区間に分割する．検索の高速化を行うため多次元インデックス構造の R* 木を使用するが，インデックスを効率良く構築・利用可能とするために，連続する区間は類似していることが多いことを利用して，複数の区間を区間群として扱う．さらに，検索精度を向上させるために，連続区間を一つの区域として扱い，かつ，区域への分割の影響を考慮して，連続する区域どうしが重複するようにし，区域単位で類似度を求める．実験により評価したところ，インデックス構築速度が速く，インデックス量も少なく，また，検索速度も速く，さらに，検索精度も良いことを明らかにした．

Fast Partial Similarity Retrieval Method of Swinging Time Series

Yoshihiro Yamauchi,[†1,†2] Teruhisa Hochin,[†1]
Hiroki Nomiya,[†1] Hideya Nakanishi[†3]
and Mamoru Kojima[†3]

This paper proposes a method of the efficient partial similarity retrieval of swinging time series. A waveform is divided into segments, which are very shorter than the waveform. A part of waveform is represented with a series of segments. The R* tree is used in order to speed up the retrieval time. Handling two or more consecutive segments as a segment group enables the index to be efficiently constructed and be used. In addition, for the purpose of improving the retrieval precision, continuous segments are treated as one "section," and the dissimilarity is calculated for each section. The adjoining district overlaps of sections could reduce the influence of dividing a waveform into sections. It is experimentally clarified that the proposed method could give us the good performance of constructing the index, the small size of the index, the good retrieval performance, and the good precision of the retrieval.

1. はじめに

核融合科学研究所等で行われている核融合科学実験では，熱量をセンサーで計測したデータやプラズマ外部の磁場の変動（磁場揺動）を計測したデータなどの，様々な，しかも，大量の測定データが得られる[1)]．これらのデータは時間とともに値が変化する時系列データであり，熱量計測データのように動きの緩やかなものもあれば，磁場揺動データのように動きの激しいものもある．これまでの実験データはストレージに蓄積されており，蓄積された大量の実験データから，最近発見された興味深い現象の実験データと類似のものを取り出したいという要求がある[2)–5)]．しかし，現状では，実験パラメータの類似性に基づいた検索が行われる程度である．実験パラメータが異なっていても波形としては類似のものがある可能性があり，また，このような波形こそが新たな発見に結びつく可能性があると考えられるのであるが，蓄積されている大量のデータの波形に対して一つ一つ目視により類似性を判定することは事実上不可能であり，計算機を利用した支援が望まれている[2)–5)]．

ここで，一般に，時系列データの検索は，大きく，全体検索と部分検索に分けられる．全体検索とは，検索対象となる時系列データの長さと検索キーとなる時系列データの長さが同じものであり，部分検索は，検索キーとなる時系列データの長さが検索対象の時系列データの長さよりも短いものである．

動きの緩やかな時系列データに対しては，時系列データ全体に対する問合せが多く[2),4),5)]，したがって，全体検索となることがほとんどである．動きの緩やかな時系列データに対しては様々な研究がなされている．最も単純な方法は，時系列データをユークリッド空間の 1 点とみなし，そのユークリッド距離を非類似度として検索を行う方法である．しかし，時系列データをユークリッド空間の 1 点とみなすと，その空間の次元は非常に高くなってしまう．多次元空間のオブジェクトを高速に求めるための多次元インデックス構造は，10 次元程度が実用的な限界であることが実験的に明らかになっており[6)]，次元を削減させるための方法が研究されている．一つの方法は，時系列データを周波数表現に変換し，そのいくつかの

†1 京都工芸繊維大学大学院 工芸科学研究科
Graduate School of Science and Technology, Kyoto Institute of Technology
†2 現在，株式会社神戸製鋼所
currently with Kobe Steel, Ltd.
†3 核融合科学研究所
National Institute for Fusion Science

Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

係数を利用する方法である[6),7)]．その他の方法として，例えば，時系列の波形をパルス波で近似し，その面積を扱うことで次元数の削減を図る APCA(Adaptive Piece-wise Constant Approximation)[6)] がある．

一方，動きの激しい揺動型の時系列データに対しては，全体検索よりもむしろ部分検索を行いたいという要求が多い[3)]．揺動型時系列データの一例として，磁場揺動データの例を**図 1** に示す．時系列データの部分検索の大きな問題は，検索キー波形の長さが固定ではなく，検索時まで決定されないことが多いことである．これに対処する最も単純な方法は，1 点から (波形全体 −1) 点までの全ての長さの部分波形に対するインデックスを構築し検索に利用することであるが，格納量や挿入性能の点から実用的とは言い難い．また，検索キーの長さが波形の長さ以下となるので，検索キー波形に類似する部分波形が，1 つの波形に多数存在する可能性も考慮しなければならない．波形の長さを s，検索キー波形の長さを q としたとき，類似波形になる可能性のある部分波形は，1 つの波形の中に $s-(q-1)$ 個存在する．例えば，$s=100$，$q=10$ の場合，部分波形は 91 個である．従って，各部分波形に対して全体検索と同様の手法を用いると，全体検索よりも高速検索のために必要なコストが確実に増大してしまう．そこで，このコストをできるだけ少なくするような工夫が必要となる．

時系列データの部分検索問題に対しては，特徴量が特徴空間に描く軌跡を切り取ってできる最小包含矩形 (Minimum Bounding Rectangle: MBR）を用いてインデックスを構築する ST（Sub-Trail）index と，検索キー系列の最初の部分だけでインデックスを探索する Prefix-Search が提案されている[8)]．また，ST-index を改良し，複数の異なる解像度のインデックスを用意する MR（Multi-Resolution）index[9)] などが提案されているが，全体検索に比べてまだまだ研究途上である．

そこで，本論文では，膨大な量の動きの激しい揺動型の時系列データから，部分時系列データを効率良く，かつ，精度良く求めることを目的として，時系列データの高速類似部分検索手法を提案する．提案手法は，波形を全体から見ると極小の長さの区間に分割する．そして，連続する区間は類似していることが多いことを利用し，単独の区間ではなく複数の区間を包含する区間群を用いることで，格納効率の向上を図る．また，検索時には，連続区間を一つの区域として扱い，かつ，区域への分割の影響を考慮して，連続する区域どうしが重複するようにし，区域単位で類似度を求めることにより検索精度の向上を図る．本論文では，実験により提案手法の評価を行う．

本論文では，まず，**2.** で関連研究と使用する非類似度について述べる．次に，**3.** で類似部分検索法を提案し，**4.** で実験により評価する．最後に，**5.** で本研究をまとめる．

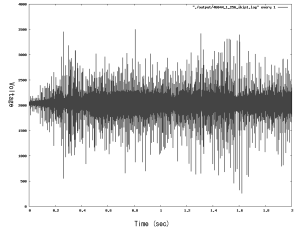

図 1 磁場揺動データの波形
Fig. 1 Waveform of magnetic field fluctuation data.

2. 時系列データの類似部分検索

2.1 関 連 研 究

Faloutsos らは，時系列データの部分検索法として， ST-index と Prefix-Search を提案している[8)]．以下，これらについて概説する．

ST-index では，最小の検索キー波形の長さ（大きさ）を l，時系列データ S の長さ（大きさ）を $Len(S)$ とすると，時系列データ上を一定の長さ u（$u<l$）ずつずらしながら，長さ l の分割波形を取り出す．この分割波形から特徴量を抽出し，特徴空間に点としてマッピングしていく．点によってできた軌跡を分割し，MBR で表現する．この MBR をインデックスに格納する．1 つの波形を複数の MBR で表現する場合，MBR の分割方法はコスト関数を使用して決定する．大きさが $L_1, L_2, \ldots, L_d$ の d 次元の MBR において，内包する分割波形の点が k 個のとき，この MBR のコスト関数 mc は式 (1) である．

$$mc = \frac{\prod_{i=1}^{d}(L_i+0.5)}{k} \tag{1}$$

新たな分割波形を既存の MBR に追加した時，コスト関数が増加したら，その MBR をインデックスに格納し，新たな MBR を作成する．MBR は以下の情報を持つ．

- 各次元における MBR の範囲 $\{(F1_{low}, F1_{high}), \ldots, (Fd_{low}, Fd_{high})\}$．ここで，$Fi_{low}$ および Fi_{high} は i 番目の次元における MBR の最小値と最大値である．
- 時系列データの識別子
- t_{start}，t_{end}：初めの分割波形と終わりの分割波形の時系列データにおける位置

Prefix-Search では，検索キー波形 Q の長さを $Len(Q)$ とすると，検索キー波形 Q の先頭から長さ l の部分波形に合う解候補を検索する．ただし $Len(Q) \geq l$ とする．後処理とし

Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

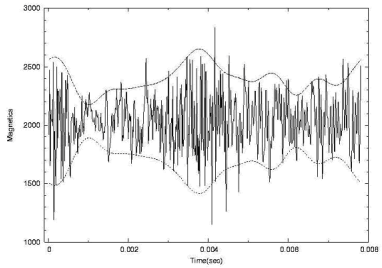

図 2 高周波成分を多く含む波形に見える外形線（破線）
Fig. 2 Outline of a waveform (short dashed line).

て解候補と検索キー系列の類似度を評価し，順位付けする．

ST-index & Prefix-Search は検索キー波形の先頭から決まった長さの検索キーの部分波形のみでインデックスを範囲検索しているので，検索漏れをなくすためには検索範囲をある程度大きくとらなければならない．そのため，検索速度が遅くなってしまう．

Kahveci らは，ST-index を改良し，複数の異なる解像度のインデックスを用意する MR-index を提案している[9]．この手法では，解像度ごとに特徴量を格納するため，インデックスサイズが増大する．インデックスの圧縮手法も提案されているが，依然として格納時間と格納量のオーバヘッドは問題である．

2.2 漸均・外形スペクトル距離

ここでは，本論文で使用する時系列データの非類似度である漸均・外形スペクトル距離について概説する．漸均・外形スペクトル距離は，高周波成分を多く持つ時系列データに対し，ヒトの視覚的認識に近い類似性の判定を目指した 2 つの特徴を持つ非類似度である[11]．1 つ目は「周波数の高低による識別性の考慮」である．ヒトの目は，低い周波数の波形の場合，周波数の違いや位相のずれなども認識できるが，周波数が高くなるにつれて，周波数や位相の違いを認識することが困難になるという特性を考慮している．2 つ目は「外形線の考慮」である．高周波成分を多く含む波形の場合，波形の上下に**図 2** の破線のような外形線が認識できる．この曲線の形が類似性の判定に有効であると考えている．

時系列データ $x = [x_t]$，$y = [y_t]$ の非類似度 $D(x, y)$ は，実数 r_1，r_2，r_3 に対し式 (2) で求められる．

$$D(x, y) = r_1 \cdot D_1(x, y) + r_2 \cdot D_2(x, y) + r_3 \cdot D_3(x, y) \tag{2}$$

ここで，整数 k,l,d 及び列 $\{i_j\}_{j=0}^{d}$ に対して以下である．

$$D_1(x, y) = \sqrt{\sum_{f=1}^{k-1} (X_f - Y_f)^2} + \sqrt{\sum_{f=k}^{l-1} (|X_f| - |Y_f|)^2} \tag{3}$$

$$D_2(x, y) = \sqrt{\sum_{f=1}^{k-1} (X'_f - Y'_f)^2} + \sqrt{\sum_{f=k}^{l-1} (|X'_f| - |Y'_f|)^2} \tag{4}$$

$$D_3(x, y) = \sqrt{\sum_{j=1}^{d} \Big(\sum_{f=i_{j-1}+1}^{i_j} (|X'_f| - |Y'_f|) \Big)^2} \tag{5}$$

$[X_f]$，$[Y_f]$ は時系列データ $[x_t]$，$[y_t]$ のフーリエ級数を表す．また，$[X'_f]$，$[Y'_f]$ は，時系列データ $[x_t]$，$[y_t]$ から l 未満の周波数成分を除いた外形線の時系列データ $[x'_t]$，$[y'_t]$ のフーリエ級数を表す．r_1，r_2，r_3 は，それぞれ，$D_1(x, y)$，$D_2(x, y)$，$D_3(x, y)$ の重みである．k は低周波と中周波を分ける閾値であり，l は中周波と高周波を分ける閾値である．d は高周波を分割する数であり，列 $\{i_j\}_{j=0}^{d}$ は分割した高周波の列であり，$i_0 = l - 1$，$i_d = n/2$（n は時系列データを構成する点の数）である．$D_1(x, y)$ は，オリジナルの時系列データの低周波成分と中周波成分の非類似度であり，$D_2(x, y)$ は，外形線の低周波成分と中周波成分の非類似度である．$D_3(x, y)$ は，高周波成分の非類似度である．

3. 提 案 手 法

3.1 概　　要

提案手法では，時系列データの非類似度として，2.2 で述べた漸均・外形スペクトル距離を採用するので，時系列データをフーリエ変換する必要がある．フーリエ変換を高速に行うために高速フーリエ変換（FFT）を利用するには，点の数は 2 のべき数とする必要がある．また，あまり多くの点を使用すると FFT といえども時間がかかってしまう．そこで，適当な 2 のべき数 w を用い，時系列データを w 点ごとの区間に区切り，区間ごとに FFT を行う．従って，時系列データを区間の列として扱うことになる．区間により時系列データを分割することになるので，時系列データの一部分を取り出すには，切り出す最初の点，および，最後の点は区間の長さの整数倍の点でなければならない．この際に，1 区間の点の数が多いと，切り出したい点で時系列データの一部分を切り出せなくなってしまう．そのため，w は時系列データ全体の長さ S に対して $w \ll S$ という条件を満足するものとする．また，w は，2.1 で述べた ST-index における最小の検索キー波形の長さ l よりも短く（$w < l$），

Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

ずらしの長さ u と同程度の長さ（$u \approx w$）とする．

膨大な時系列データに対する類似検索では，その検索コストが問題となる．本研究で対象としている磁場揺動の波形も同様であり，この問題を解決するために，R*木[10] を用いて多次元インデックスを構築する．R*木を使用したのは，R*木が多次元データに対する高速検索を可能とするインデックス構造である R 木の改良版であり[10]，最も良く利用されているものの一つであるからである．

ここで，漸均・外形スペクトル距離を求めるためのフーリエ係数を特徴量と呼ぶ．特徴量の数は通常 100 個以上となる[11]．これをこのまま R*木に格納するということは，特徴量を 100 次元以上の空間中の点として扱うということであり，多次元インデックスにおける次元の呪い[6] の問題から好ましくない．そこで，特徴量から代表となるもの（以降，代表特徴量と呼ぶ）を抽出して R*木に格納する．これについては 3.2 で詳述する．一方，漸均・外形スペクトル距離を求めるためにはすべての特徴量が必要である．すべての特徴量を R*木に属性として格納することも考えられるが，R*木が大きくなってしまい，検索効率が悪くなってしまう恐れがある．また，検索時に再度 FFT を行う方法も考えられるが，検索時の性能が悪くなってしまう．さらに，特徴量を DB 中のテーブルとして保持することも考えられるが，DB 管理システムのオーバヘッドが見積もりにくいという問題がある．そこで，ここでは特徴量をバイナリ形式でファイル（以降，特徴量バイナリファイルと呼ぶ）に格納することとする．

多次元インデックスを用いた部分検索は，大きく分けてインデックス構築フェーズと検索フェーズの二つのフェーズから成る．各フェーズの動作の大まかな流れを以下に示す．

【インデックス構築フェーズ】

(1) 波形を長さ w で区切り，区間とする．
(2) 各区間より特徴量を抽出し，特徴量バイナリファイルに格納する．
(3) 特徴量から代表特徴量を選出し，多次元インデックスを構築する．

【検索フェーズ】

(1) 検索キー波形を区間に分割し，特徴量を抽出する．
(2) その特徴量から代表特徴量を選出し，代表特徴量をキーとして多次元インデックスに対して検索を行う．
(3) 得られた波形の集合（解候補集合）をもとにして特徴量バイナリファイルより各波形の特徴量を取得する．
(4) 検索キー波形の特徴量と，(3) で得た特徴量を用いて非類似度を算出する．

3.2 特徴量および代表特徴量

ここでは，漸均・外形スペクトル距離を求める際のパラメタ（2.2 で示した式 (2) から式 (5) のパラメタ）を，文献[11] に基づき，それぞれ，$r_1 = 1$，$r_2 = 5$，$r_3 = 6$，$k = 7$，$l = 57$，$d = 8$ とする．この時の特徴量の数は 133 である．また，特徴量の各帯域から周波数が低い 2 点を選出し，計 15 点を代表特徴量とする（ただし，低周波域と外形線低周波域は例外）．これらの内訳をまとめて**表 1** に示す．

全特徴量は，波形の識別番号，区間番号の順に特徴量バイナリファイルに格納し，インデックス検索後の非類似度計算を行う際に使用する．

3.3 区間群インデックス

区間番号が近い区間は視覚的に似ているものが続く傾向にある．そこで，同じ波形内において，連続して似ている区間を 1 つの区間群としてまとめることでインデックスのデータサイズを削減する．

ここで，区間群を 15 次元空間中の一点として表現した場合を考えると，似ている区間はより近い場所に集まっているはずである．そこで，連続して似ている区間の点の集まりを，それらを内包する MBR としてインデックスに格納する．ここで，代表特徴量は 15 点なので， 次元数 $d = 15$ である．

一つの波形を複数の MBR で表現するので，2.1 で述べたコスト関数（式 (1)）を利用して MBR を決定する．この方法で構築したインデックスを区間群インデックスと呼ぶ．

3.4 区域検索法

区間の長さを w，区間の数を n とし，検索キー波形の長さを $w \times n$ とする．まず，この検索キー波形を用いてインデックスを引くことを考える．この場合，検索キー波形を長さ w で等分割し，できた n 個の区間の代表特徴量でインデックスを引き，その結果を解候補集合とする方法が一番単純である．しかし，この方法では n が大きい場合，インデックスを

表 1 特徴量と代表特徴量の内訳
Table 1 Breakdown of features and representative features.

帯域	特徴量の数	特徴量の内容	代表特徴量の数	代表特徴量の内容
低周波	12	原点含まず，実部 6，虚部 6	4	原点含まず，実部 2，虚部 2
中周波	50	成分の大きさ	2	成分の大きさ
外形線 低周波	13	原点含み，実部 6，虚部 6	5	原点含み，実部 2，虚部 2
外形線 中周波	50	成分の大きさ	2	成分の大きさ
外形線 高周波	8	成分の大きさ	2	成分の大きさ
計	133		15	

Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

表 2 提案手法と比較手法の比較
Table 2 Comparison of the proposed and the comparative methods.

手法	インデックス	検索方法	非類似度	非類似度の計算方法
提案手法	区間群インデックス	区域検索法	漸均・外形スペクトル距離	区域ごと
比較手法	ST-index	Prefix-Search	同上	区間ごと

引く回数も n 回と多くなるので，検索コストが大きくなってしまう恐れがある．

次に，検索キー波形と解候補の各波形とで非類似度を求めることを考える．この場合，検索キー波形と解候補の各波形を共に長さ w で等分割し，対応する区間ごとに非類似度を求め，それを合算する方法が考えられる．しかし，n が大きい場合，検索キーと比較して相対的に極めて小さい区間ごとの比較となるため，視覚的に似ているにもかかわらず，区間がずれているために非類似度が大きくなってしまう恐れがある．

これらの問題を解決するために，インデックスを引くとき，および，非類似度を計算するときに，区間単位で行うのではなく，複数区間での各特徴量の平均値を使用することとする．以後，この区間を複数集めたものを区域と呼ぶ．また，多少のずれを吸収するために，区域どうしのいくつかの区間を重複させる．この方法を区域検索法と呼ぶ．

4. 評　　価

4.1 評 価 方 法

前述の提案手法を実装し， ST-index & Prefix-Search 手法と比較して評価を行った．比較手法では，インデックスに ST-index を用い，インデックスの近傍探索に Prefix-Search を用いた．比較手法は，本来，非類似度にはユークリッド距離を用いるが，このままでは比較にならないほど検索精度が悪いので，漸均・外形スペクトル距離を用いた．また，インデックスを引いて出た解候補と検索キー波形での非類似度の算出は，区間ごとに行う従来の方法を用いた．提案手法と比較手法をまとめて**表 2** に示す．また，波形の長さ $S = 131072$，部分波形の最小長 $l = 4096$，区間の長さ $w = 512$，区域の長さは 8192（16 区間），ずらし長は，提案手法が 4096（8 区間）で比較手法が 512 である．測定には，カスタム PC（Intel Celeron M(1.4GHz)，1GB メモリ，Serial ATA HDD(250GB, 7200rpm)）に Vine Linux3.2 をインストールして使用した．また，R*木インデックスを構築するプログラムは，Hadjieleftheriou 氏作成のものを使用した[12)⋆1]．

⋆1 使用したプログラムのバージョンは 1.3.0.

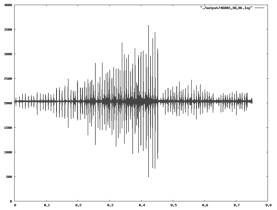

図 3 検索キー波形の一例
Fig. 3 An example of a key waveform.

4.1.1 実験データ

測定データは，1 波形あたり 131,072 点で構成されている．本実験ではこの波形を最大 2000 個用いた．各波形を 512 点つまり，$w = 512$ ごとに区切り，256 個の区間を作成した．こうしてできた 2000 × 256 個の区間を実験データとした．

4.1.2 インデックス構築速度・インデックスサイズ評価方法

波形をそれぞれ，1000 個，1500 個，2000 個用いて，区間群インデックスと ST-index の構築を行い，インデックス内のデータ（MBR）数とインデックスのサイズ，ならびに，構築速度を計測した．なお，構築時間を測定する際は，メインメモリ上に関連するデータが無い状態（cold 状態）で行った．

4.1.3 検索速度評価方法

検索キー波形の長さを 64,96,128,160,192 区間（波形全体は 256 区間）とし，キー波形の長さを変えた時の検索時間の変化を計測した．検索方法としては，近傍検索を用いた．また，両手法ともインデックスを引いて求める解候補の数は約 7000 個となるようにパラメタ（N 近傍検索の N の値）の調節を行った．**図 3** は検索キー波形の一例で，長さが 96 区間のものである．なお，検索速度を測定する際は，cold 状態とメインメモリ上に関連するデータが在る状態（hot 状態）の 2 種類の状態を用意して行った．

4.1.4 検索精度評価方法

視覚的類似性の高い波形が検索できているかを検証するための評価実験を行った．

図 4 と**図 5** は，実験で用いる検索キー波形であり，どちらも長さは 192 区間である．図 4 は比較的低周波を多く含む波形であり，図 5 は高周波を多く含む波形である．まず，著者の一人が検索キーの波形と検索対象となる波形を比較し，視覚的に似ている波形の選定（正解集合の選定）を行った．この実験では波形 1000 個を用意し，1 波形に対し，区間を 8 個ず

Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

つずらして 8 枚の波形を出力した．従って，計 8000 枚の波形から正解集合の選定を行った．図 4 と図 5 に示した検索キー波形に対する正解波形の数は，おのおの，147 と 166 である．

評価には適合率と再現率を用いた．適合率＝検索された正解集合の要素数／検索総数であり，再現率＝検索された正解集合の要素数／正解集合全体の要素数である．

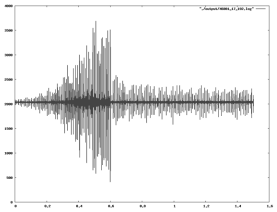

図 4 検索精度評価実験キー波形 1
Fig. 4 key1 used in similarity evaluation experiment.

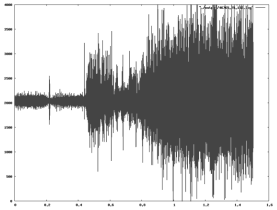

図 5 検索精度評価実験キー波形 2
Fig. 5 key2 used in similarity evaluation experiment.

4.2 実験結果と考察

4.2.1 インデックス構築速度・インデックスサイズ評価

区間群インデックス（SG-index）と ST-index の構築時間を**図 6** に示す．図中の user time は主に数値計算に要した CPU 時間であり，system time は主に入出力に要した時間である．

図より区間群インデックスのほうが ST-index よりも約 4 倍高速である．system time には大きな差はなく，user time の差が大きくなっており，ST-index では数値計算に使用する CPU 時間が多くかかっているということである．ST-index では，1 つの波形を 512 点ずつずらしながら，4096 点ごとに 250 個の分割波形を作成している．一方，区間群インデックスでは，512 点ごとに分割して 256 個の区間を作成する．分割波形の個数と区間の個数に大きな差がないことから，user time の差は，1 つの分割波形を FFT するためにかかる時間の差の合計であると考えられる．

ST-index において分割波形の長さを小さくすれば，1 つの分割波形を FFT するためにかかる時間は少なくなるが，その分個数が多くなってしまう．また，ST-index は分割波形どうしの重なり部分が生じるため，例え分割波形の長さを 512 点としても個数は区間群インデックスよりも多くなってしまう（ST-index において分割波形の長さが 512 点で，ずらしの大きさが分割波形の長さと等しいとき，区間群インデックスと同等となる）．

次に，インデックスのデータ（MBR）数と，インデックスサイズをまとめて**図 7** に示す．図 7 では，インデックスのデータ数を棒グラフで示し，インデックスサイズを折れ線グラフで示している．インデックスのデータ数では，区間を区間群にまとめずそのままインデックスに格納した場合も“ Segments”として示している．インデックスに格納したデータ数では，区間群インデックスは，区間をそのままインデックスに格納した場合よりも約 75 %，ST-index よりも約 20 %データ数が減少しており，データ格納コストの削減に成功していることが分かる．また，インデックスサイズでは，区間群インデックスは ST-index の 78 %のサイズであり，インデックスサイズを低減できている．

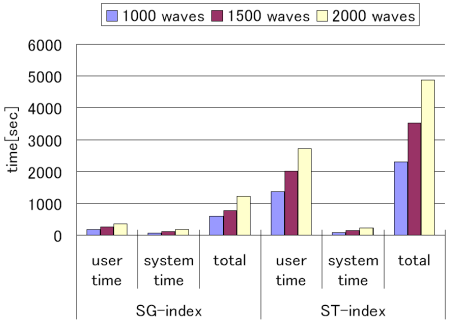

図 6 インデックス構築性能評価結果
Fig. 6 Time of creating indexes.

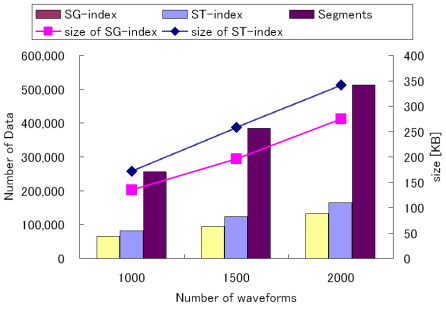

図 7 インデックスのデータ数とインデックスサイズ
Fig. 7 Numbers of data and index sizes.

4.2.2 検索速度評価

cold 状態の検索速度を**図 8**，hot 状態の検索速度を**図 9** に示す．これらの図では，提案手法を「proposed」，比較手法を「ST&Prefix」と記している．

区間数が 32 から 128 までは，所要時間が線形に増加する傾向が見られる．しかし，160 以上の区間数での所要時間は，この傾向からはずれている．これは，1 波形の区間数が 256 であるため，区間数が 128 を超える場合はこの区間数を持つ部分波形が少なくなってしまうためと考えられる．

図から，cold 状態では，区間数が 32，64 の場合は区域検索法と Prefix-Search に速度差はほぼ無いが，区間数が大きくなるにつれて差が広がり，区間数が 192 において最大となる．このとき区域検索法の方が Prefix-Search よりも 30 %ほど高速である．hot 状態では区域検索法が Prefix-Search よりも 40 %から 80 %ほど高速である．また，キー波形が大き

Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

くなればなるほど，検索性能差が大きくなる傾向にある．これは，インデックスの大きさの違い，および，非類似度計算を区間で行うか，区域で行うかの差が出ているためであると考えられる．

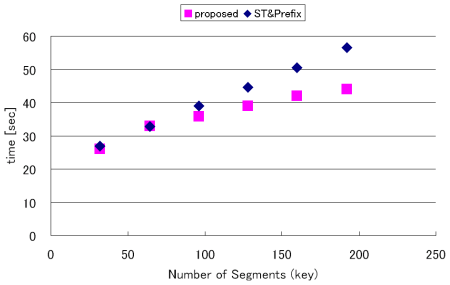

図 8 検索速度評価結果（cold 状態）
Fig. 8 Times of retrieving waveforms (cold).

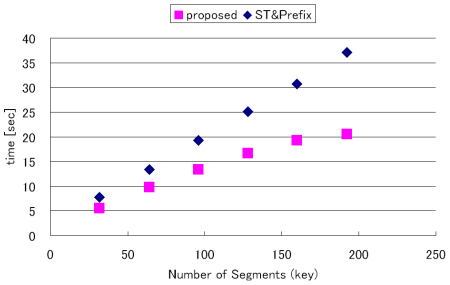

図 9 検索速度評価結果（hot 状態）
Fig. 9 Times of retrieving waveforms (hot).

4.2.3 検索精度評価

検索キー波形 1（図 4）と検索キー波形 2（図 5）に対する再現率・適合率曲線を，それぞれ，**図 10** と**図 11** に示す．

検索キー波形 1 については，ST-index & Prefix-Search では再現率が 0.2 あたりで適合率が急激に低下しているのに対して，提案手法では再現率が 0.7 あたりまで適合率がほとんど低下していない．また，検索キー波形 2 については，ST-index & Prefix-Search では再現率が 0.1 あたりで適合率が急激に低下しているのに対して，提案手法では再現率が 0.3 あたりまで適合率が低下していない．

提案手法の方が検索精度が良いのは，区域を重複させて類似度を計算しており，多少のずれをうまく吸収できているからであろうと考えられる．また，図 11 の方が図 10 よりも検索精度が悪いのは，検索波形の高周波成分を多く含むためであろうと考えられる．これは，著者らの先行研究の結果[11] と合致している．

以上より，提案手法を用いた方が検索精度が良いといえる．

5. おわりに

本論文では，多次元インデックスを用いて，動きの激しい揺動型の時系列データを対象に，波形の一部分をキーとして高速に類似検索を行う手法を提案した．提案手法では，波形

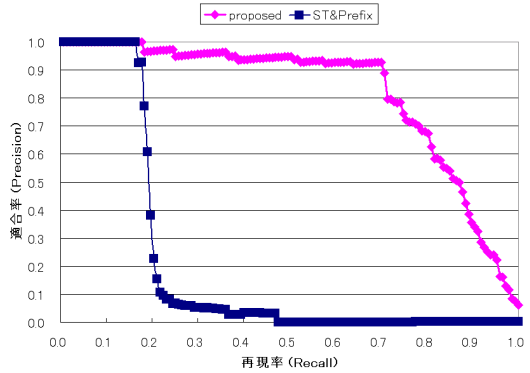

図 10 検索精度評価結果 1
Fig. 10 Recall-Precision curves for key1.

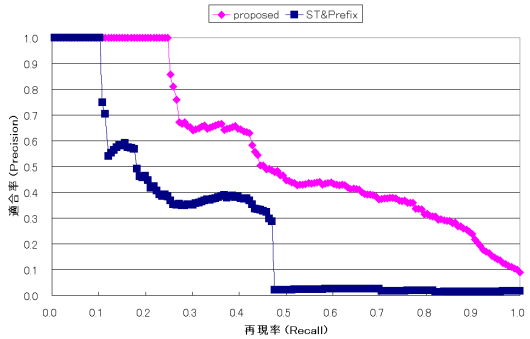

図 11 検索精度評価結果 2
Fig. 11 Recall-Precision curves for key2.

を全体から見ると極小の長さの区間に分割した．多数の解像度でのインデックスを構築し検索に利用する方法と比較して，提案手法では，区間という一つの解像度のみでインデックスを構築しており，インデックスの容量を少なく抑えているといえる．しかし，波形を細かく分割し，検索時にそれらをまとめて類似性を判定しようとすると，多数回インデックスを引かなければならず，また，区間が少しずれただけでも非類似度が大きくなってしまう．そこで，提案手法では，連続区間を一つの区域として扱い，かつ，連続する区域どうしが重複す

Vol.2009-DBS-148 No.17
Vol.2009-FI-95 No.17
2009/7/28

るようにし，区域単位で類似度を求めることでこの問題を克服した．そして，実験により，既存の検索手法よりも良い性能を示すことを明らかにした．

本論文では，時系列データの非類似度として漸均・外形スペクトル距離を使用した．これは，揺動型の時系列データに対しては，ユークリッド距離では適正な非類似度を計算できないからである．この問題は，ダイナミックタイムワーピング[13] やメルケプストラム[14] を使用することで対処することも考えられる．ダイナミックタイムワーピングを用いる方法とは，２つの時系列データの距離を最小化するように時間軸を伸長させる変換処理を行い，これによって得られた距離を時系列データの非類似度とするものである[13]．ダイナミックタイムワーピングでは，長さの異なる時系列データを扱うことが可能である．しかし，検索時の類似度計算の計算量が大きくならざるをえないという問題がある．また，メルケプストラムは時系列データをフーリエ変換し，その係数の絶対値の対数を，人間の聴覚特性（低い周波数では細かく，高い周波数では荒い分解能を持つ）に合わせたフィルタ群を使ってまとめ，それを逆フーリエ変換したものである[14]．メルケプストラムでは，低周波帯での位相のずれを考慮できず，視覚的な特徴をうまく捉えられない[15]．これに対して，本論文で使用した漸均・外形スペクトル距離は，時系列データの周波数帯を低周波域，中周波域，高周波域の３つに分類して評価し，また，時系列データの外形線を考慮しており，揺動型の時系列データに対しても精度の良い非類似度となっている．

提案手法は，検索の際に区域の数だけインデックスを検索している．類似の区域についてはインデックス検索を行わない等によりインデックス検索の回数を減らす工夫を行うことで，検索速度を向上させることができるのではないかと考えている．また，精度評価において，検索キー波形の長さを変えての実験の実施，他のタイプの波形データへの適応，より多くの波形データを用いての性能評価なども今後の課題である．

謝辞　本研究は，一部，NIFS 共同研究 NIFS06KCHH009 により支援されている．

参 考 文 献

1) 核融合科学研究所大型ヘリカル装置情報：http://www.lhd.nifs.ac.jp/.
2) Nakanishi, H., Hochin, T., Kojima, M. and LABCOM group: Search and retrieval method of similar plasma waveforms. *Fusion Engineering and Design*, Vol. 71, No. 1-4, pp.189–193 (2004).
3) Nakanishi, H., Hochin, T., Kojima, M. and LABCOM group: Similar pattern search for time-sectional oscillation in huge plasma waveform database, *Fusion Engineering and Design*, Vol.81, pp.2003–2007 (2006).
4) Dormido-Canto, S., Vega, J., Sanchez, J. and Farias, G.: Information retrieval and classification with wavelets and support vector machines. *Lecture Notes in Computer Science*, Vol. 3562, pp.548–557 (2005).
5) Farias, G., Dormido-Canto, S., Vega, J., Sanchez, J., Duro, N., Dormido, R., Ochando, M., Santos, M. and Pajares, G.: Searching for patterns in TJ-II time evolution signals. *Fusion Engineering and Design*, Vol. 81, pp.1993–1997 (2006).
6) Keogh, E., Chakrabarti, K., Mehrotra, S. and Pazzani, M.: Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases, *Proc. of ACM SIGMOD2001*, pp.151–162 (2001).
7) Agrawal, R., Faloutsos, C. and Swami, A.N.: Efficient Similarity Search In Sequence Databases, *Proc. of 1993 Int'l Conf on Data Organization (FODO'98)*, pp. 69–84 (1998).
8) Faloutsos, C., Ranganathan, M. and Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases, *Proc. ACM SIGMOD Conf.*, pp.419–429 (1994).
9) Kahveci, T. and Singh, A.: Optimizing Similarity Search for Arbitrary Length Time Series Queries, *IEEE Transactions on Knowledge and Data Engineering*, Vol.16, No.4, pp.418–433 (2004).
10) Beckmann, N., Kriegel, H.-P., Schneider, R.and Seeger, B.: The R*-tree: an efficient and robust access method for points and rectangles, *Proc. of the 1990 ACM SIGMOD Conf.*, pp. 322–331 (1990).
11) 小山克正，宝珍輝尚，中西秀哉，小嶋　護：時系列データの周波数に基づく類似度について，情報処理学会第 85 回情報学基礎研究会，pp.53–60 (2006).
12) Hadjieleftheriou, M.: http://www.research.att.com/marioh/index.
13) 大桃　諭，陳　漢雄，古瀬一隆，大保信夫：タイムワーピングに基づく時系列データの類似検索次元縮小による効率化，*DBSJ Letters*, Vol.4, No.1, pp.17–20 (2005).
14) 荒木雅弘：フリーソフトで作る音声認識システム，森北出版 (2007).
15) 山内祥裕，宝珍輝尚：周波数に基づく波形の非類似度について，平成 20 年度情報処理学会関西支部大会講演論文集，pp.113–116 (2008).