データベースシステム 128－71
（２００２．７．１９）

波形特徴を用いた類似シーケンス検索

川島 英之† 遠山 元道§
今井 倫太§ 安西 祐一郎§

†慶應義塾大学大学院 理工学研究科 開放環境科学専攻
§慶應義塾大学 理工学部 情報工学科
E-mail: kawasima@ayu.ics.keio.ac.jp, toyama@ics.keio.ac.jp,
michita@ayu.ics.keio.ac.jp, anzai@ayu.ics.keio.ac.jp

センサデータストリームから状況を認識し、その状況に応じて処理を行うアプリケーションにとって、類似シーケンスの高速検索は必須である。本研究では、シーケンスに対して、窓内に存在する山の数と面積をキーとして空間索引を張ることで、類似シーケンスの検索速度を向上させる。提案手法の検索精度と検索速度を実験により力任せ法と比較したところ、提案手法の精度は力任せ法より劣るが、検索速度は力任せ法よりも最大で約 778 倍になる結果が得られた。

キーワード：波形特徴 、 シーケンス検索 、 空間索引 、 データストリーム

Retrieval of Similar Sequences with Waveform Features

Hideyuki KAWASHIMA† Motomichi TOYAMA§
Michita IMAI§ Yuichiro ANZAI§

†School of Science for OPEN and Environmental Systems,
Faculty of Science and Technology, Keio University.
§Department of Information and Computer Science,
Faculty of Science and Technology, Keio University.
E-mail : kawasima@ayu.ics.keio.ac.jp, toyama@ics.keio.ac.jp
michita@ayu.ics.keio.ac.jp, anzai@ayu.ics.keio.ac.jp

Recently new kinds of applications that monitor sensor data streams and behave depending on contexts have appeared. The retrieval of similar sequences is essential for them to recognize current context in real-time. In this paper we propose a new similar sequcence indexing technique with waveform features. Our technique further utilize a spatial index. We evaluated the proposed method about accuracy and speed. The results showed that the proposed method could provide some degrre of accuracy, and was better about speed.

Keyword : Waveform Feature , Similar Sequence Retrieval , Spatial Index , Data Stream

1 はじめに

近年、移動体位置管理 [1]、オンライン金融解析、そしてロボット知的処理 [2] などのアプリケーションなどが現れてきている。これらのアプリケーションは、途切れなく到着するセンサデータストリーム群を監視することにより現在の状況を取得し続けるとともに、状況に合わせた行動をする必要がある。移動体位置管理システムは変わり続ける移動体の位置を取得し続けるだけでなく、目的地までの最短経路情報など、付加的なサービスを提供すべきである。オンライン金融解析システムは変動する金融データをもとにして、数学的手法や AI 的手法を用いて現在の市場動向を解析し続ける必要がある。そしてロボットが人間と円滑にインタラクションをするには、リアルタイムに人間の意図を認識し、さらにその場にふさわしい振る舞いをする必要がある。

従って、これらのアプリケーションにとって、センサデータストリームからの状況認識は必須である。状況認識には様々な技術があるが、その中のひとつに事例ベースマッチング [3] がある。この技術はあらかじめ膨大な事例をデータベースに蓄えておき、問い合わせされたクエリシーケンスと類似した事例を検索し、過去の状況から現在の状況を推定する。この技術を前述のようなアプリケーションに適用すると、膨大なセンサデータシーケンスをあらかじめデータベースに蓄えておき、途切れなく到着するセンサデータストリームとデータベース中のシーケンスをマッチングし続けることになる。この認識にはリアルタイム性が要求されるから、検索処理は高速化されるべきである。

そこで本研究では、シーケンスの波形情報を利用して類似シーケンスを高速に検索する手法を提案する。提案手法は、まずシーケンスを APCA と呼ばれる形式に変換して抽象化した後、それらに対して一定サイズの窓を掛け、窓内に存在する波の数とそれらの面積の合計を計算し、そのふたつのパラメータを次元として空間索引のひとつである SR-tree により索引付けする。提案手法は力任せ法より精度が落ちるが、大きな速度改善を果たす。

本論文の構成は次の通りである。2 節ではシーケンス検索に関する従来研究について述べる。3 節では提案手法を述べる。4 節では提案手法を速度と精度の両面から評価した結果を述べる。5 節では提案手法に対する議論をおこなう。最後に 6 節で本研究をまとめる。

2 対象シーケンスと従来研究

2.1 対象データ

我々はロボットに人間とコミュニケーションさせるためのデータベースシステムを開発している。ロボットが人間とコミュニケーションをするには、(1) 現在の状況を認識し、(2) 状況にふさわしい振る舞いをする必要がある。状況認識をさせるための方針として、データベースに膨大な状況テンプレートを蓄えておき、それらの状況テンプレートとセンサデータから得られる現在の状況を比較することを考えている。ここで状況テンプレートは加工されない状態のセンサデータにより構成される。そして、コミュニケーションはリアルタイムにおこなわれるべきだから、状況テンプレートの検索は高速処理されるべきである。状況テンプレートはシーケンスとして表現されるから、類似シーケンス検索の高速化が必要である。

2.2 従来手法

類似シーケンス検索には様々なアプローチが取られてきた。本節ではそれらについて述べる。

- 力任せ法

  力任せ法では、クエリシーケンスがデータベース中の全てのシーケンスを走査する。走査は次のように行われる。クエリシーケンスを Q、Q の長さを n、そしてデータベース中のシーケンスを D とすると、力任せ法は次のように行われる。

  1. Q の先頭を S の先頭に設置
  2. Q が D の末尾に移動するまでステップ 3 と 4 を繰り返し
  3. 次式により Q と D の部分との差を計算

  $$\sum_{i=0}^{n} \sqrt{(D_i - Q_i)^2} \qquad (1)$$

  4. Q を後方にシフト

力任せ法はシーケンス全体との比較をおこなうために、最も類似した部分シーケンスを発見できる。しかしその一方で、上記の通り膨大な計算量が必要であるために、検索速度は遅い。

- 空間索引の利用

  力任せ法と同程度の精度を保証しながら検索時間を短縮するために、空間索引が使われてきた。空間索引は N 個のパラメータを N 次元空間中の点として表現し、索引を張る技術である。空間索引には R*-tree[4] や SR-tree[5] など、多くの種類がある。シーケンスに空間索引を適用する場合には、シーケンスを構成する各点を空間索引中の一次元とみなし、シーケンス長が N であれば、空間索引の次元数は N となる。

  次元数が低い際には空間索引は有効だが、次元数が高くなるにつれてパフォーマンスが劣化する。次元数が 10 を超えてしまうと、多くの空間索引では、空間索引による検索速度が力任せ法による検索速度よりも劣る事実が知られている [6][7]。

- 次元削減と空間索引の組合せ

  高次元空間において空間索引のパフォーマンスを高めるために、空間索引を張る前に、シーケンス長を減らすことで次元を削減する手法が提案されてきた。よく知られている次元削減手法には、離散フーリエ変換、離散ウェーブレット変換、移動平均、そして APCA(Adaptive Piecewise Constant Approximation) がある [7]。離散フーリエ変換は定常波でシーケンスを表現することで次元を減らす。離散ウェーブレット変換は非定常波でシーケンスを表現することで次元を減らす。移動平均は固定長の窓をシーケンスに掛け、窓内の平均値でシーケンスを表現することで次元を減らす。APCA は可変長の窓をシーケンスに掛け、窓内の平均値でシーケンスを表現することで次元を減らす。

  次元削減の問題点は、精度が落ち得ることだ。は似ているふたつのシーケンスが、次元削減の結果、類似しなくなる可能性がある。例えばパルス状のシーケンスに対する次元削減として、離散フーリエ変換とハール基底を用いた離散ウェーブレット変換を適用することは好ましくないことが知られている [7]。

3 提案

前節において、シーケンス検索には、次元削減と空間索引の組合せが好ましい事を述べた。本研究で対象とするのはパルス状のシーケンスであるために、離散フーリエ変換とハール基底ウェーブレット変換の適用は好ましくない。一方 APCA はパルス状のシーケンスを的確に表現できる。そこで本研究では APCA 表現を利用する。APCA は可変長窓をシーケンスに掛けて平均値を計算し、その平均値を用いてシーケンスを表現する。窓が可変長であるために空間索引を適用することは難しく、論文 [7] では APCA への R-tree の適用手法が新たに提案されている。

本研究では APCA に対して直接の空間索引は張らない。その代わり、APCA 表現されたシーケンスの抽象度を高めた後に、空間索引を適用する。以後、APCA 表現されたシーケンスを APCA 表現シーケンスと呼ぶ。提案手法のアプローチは、APCA 表現シーケンスに対して固定長の窓を掛け、窓内に存在する山の数および山々の面積合計をキーとして二次元の空間索引を張ることである。この手法について、本章の残りで述べる。

3.1 索引作成

1. APCA 表現

   索引作成の第一段階はシーケンスを APCA 表現に変換することである。この処理により、シーケンス中に存在するパルス状の山の特徴を消さずに、なだらかな部分の細かい山を消去する。図 1にオリジナルシーケンス、移動平均 (Mean Average) により表現されたシーケンス、APCA 表現シーケンスを重ね たグラフを示す。図 1は、APCA 表現がオリジナルの特徴を残せることを示している。一方、移動平均により表現されたシーケンスはオリジナルシーケンスのもつパルス状の特徴を大きく損なうことを示している。

2. 線分近似表現

   図 2に示されている、二つのシーケンスは類似

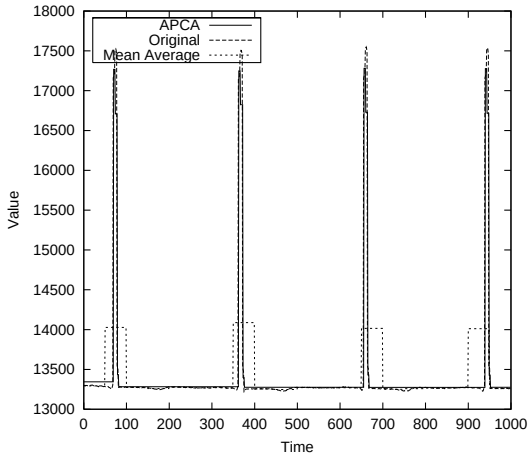

図 1: オリジナルシーケンス、APCA 表現シーケンス、移動平均シーケンスの比較

しているが、左図のシーケンスに含まれる山の数は、右図のシーケンスに含まれる山の数よりもかなり多い。提案手法のアプローチは、山の数と面積による索引付けであるから、このような種類の山は提案アプローチの検索精度を劣化させ得る。そこで本研究では隣接するパルス間の距離が小さく、かつそれらのパルスの高さの差も小さい場合には、二つのパルスを結合する。これを線分近似表現シーケンスと呼ぶ。

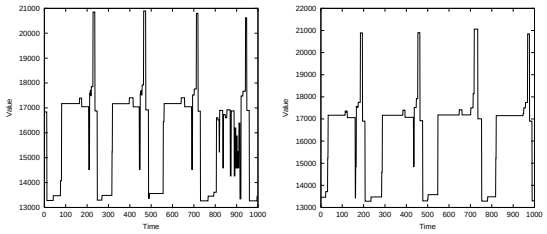

図 2: APCA の問題点

3. 山の同定

次に線分近似表現シーケンスから山を切り出す。全ての山は上に凸となるように切り出す。シーケンスの左端の点に注目する。注目点よりもその右隣接点が大きく、かつその隣接点のさらに一つ右隣接点が閾値以上に大きければ、注目点を山の始まりと同定する。山の始まりを同定した後、注目点を右に動かしていき、もしも注目点の右隣接点が注目点よりも小さければ、注目点を山の頂点と同定する。山の頂点を同定した後、注目点を右に動かしていき、もしも注目点と右隣接点の値が同じか、もしくは注目点よりも右隣接点が大きければ、注目点を山の終わりと同定する。

4. 空間索引

固定サイズの窓を線分近似表現シーケンスに掛けていき、窓内の山について、山の個数と山の総面積を計算する。この二つのパラメータをキーとして、空間索引により索引を張る。窓掛け処理の始点は山の終わりと山の始まりの中間に設定する。

3.2 検索処理

検索は次のように処理される。

1. 問い合わせの受理

問い合わせシーケンスを受け取る。ここで問い合わせのサイズは索引作成時に用いた窓と同サイズに限定する。

2. APCA 表現

問い合わせシーケンスを APCA 表現に変換する。索引作成の部分で述べた処理と同様の処理をおこなう。

3. 線分近似表現

APCA 表現シーケンスを線分近似表現に変換する。索引作成の部分で述べた処理と同様の処理をおこなう。

4. 窓掛け

線分近似表現された問い合わせシーケンスに含まれる山の数および、それらの山の面積の合計を計算する。索引作成の部分で述べた処理と同様の処理をおこなう。

5. 最近傍探索

問い合わせシーケンスに含まれる山の数および山の面積合計をキーとして、最近傍探索をおこなう。

4 評価

提案手法を評価するために実験システムを構築し、実験をおこなった。本節では実験システムの実装および実験について述べる。

4.1 実装

実験システムの実装にはC言語を用いた。ハードウェアにはSunBlade 1000を用いた。CPUは900メガヘルツのSun UltraSPARC-IIIをふたつ搭載している。実メモリのサイズは2ギガバイト、仮想メモリのサイズは5.3ギガバイト、オペレーティングシステムはSolaris 2.8である。

空間索引の実装には、SR-treeライブラリのバージョン1.3.1[8]を利用した。SR-tree[5]は包囲球と包囲矩形の両方を用いた空間索引技術である。

4.2 実験

実験データは[9]から取得した。長さが65万点であるシーケンスをすべてメモリ上に格納し、検索処理に要した時間を測定するとともに、検索処理の結果得られたシーケンスを問い合わせシーケンスと比較した。問い合わせシーケンスの長さは1000点にした。

力任せ法は比較する度に窓をずらすが、この移動幅は10とした。この幅を広くすれば検索処理が速くなるが精度が落ちる。一方、この幅を狭くすれば精度が高まるが検索処理が遅くなる。また、力任せ法は全てのシーケンスを走査した後に、精度が高い順番に整列処理をした。

実験結果について、以下に述べる。

- 速度比較

  それぞれ三回の実験を行い、その平均を取った結果を図3に示す。図3によりシーケンスが長くなるに従って力任せ法による検索速度が急激に劣化することが示される。一方、提案手法の検索速度はほとんど変化しないも示される。この理由は、提案手法では空間索引を使っているために、探索コストが小さくなるからだと推定される。提案手法は力任せ法に比べてかなり速かった。シーケンスの長さが65万点であるとき、提案手法は力任せ法に比べて、および778倍の速度差が示された。

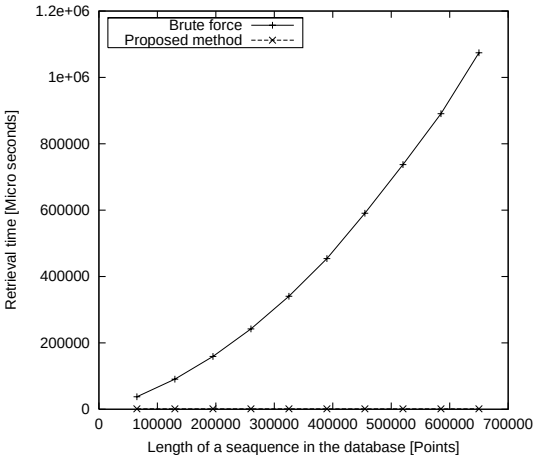

図 3: 力任せ法と提案手法の検索速度比較

- 精度比較

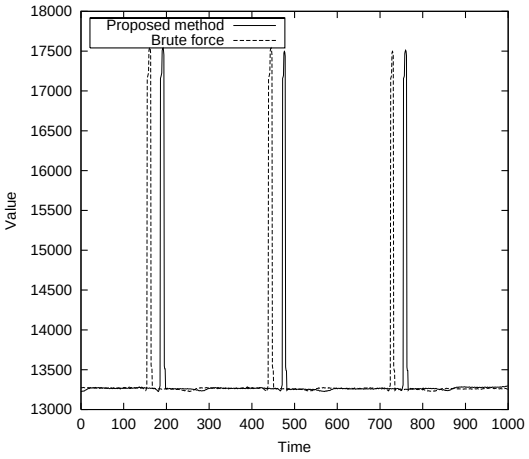

図 4: 力任せ法と提案手法の精度比較

検索により得られたシーケンスと、クエリに用いたシーケンスを図4に示す。力任せ法も提案手法も、クエリシーケンスにもっとも類似したシーケンスを表示している。力任せ法ではクエリシーケンスと全く同じシーケンスを検索することができた。提案手法ではクエリシーケンスと同じではないが、見た目がかなり類似したデータを検索することができた。

実験結果より、提案手法の検索速度は力任せ法よりもかなり優れ、さらにクエリシーケンスとある程度類似したシーケンスを検索できることが実験により示された。

5 議論

- パラメータ選択

  本研究では空間索引のキーとして、窓に含まれる山の数と、それらの面積の合計を用いた。本実験の結果ではそれなりの精度が得られたが、検索精度は高められることが好ましい。一般に、次元数が 8 程度ならば次元の呪いは発生しないと言われている。本実験では二次元しかキーを用いなかったので、今後、新たな索引キーが追加されるべきである。

- 多次元シーケンスの索引

  本研究では一次元シーケンスに対する索引手法のみ述べた。しかし、ロボットやオンライン金融解析システムといったアプリケーションにとっては、複数シーケンスに対する索引付けがされることが好ましい。なぜなら、ロボットは画像、音声、触覚など複数のセンサを同時に処理するし、オンライン金融解析システムは株式市場の解析時に複数の株価シーケンスを処理するからである。

6 結論

本研究ではセンサデータストリームを処理するアプリケーションのために、一次元シーケンスに対して、窓に含まれる山の数と面積をキーとして空間索引を張ることで、類似シーケンスの検索速度を向上させた。提案手法の検索精度と検索速度を実験により力任せ法と比較したところ、精度については力任せ法より劣るものの、検索速度は力任せ法よりも最大で約 778 倍速くなる結果が得られた。

参考文献

[1] Michimune Kohno and Yuichiro Anzai. An Adaptive Sensor Network System for Complex Environments. In *Proceedings of 5th International Conference on Intelligent Autonomous Systems*, pp. 21–28, 1998.

[2] Yuichiro Anzai. Human-Robot-Computer Interaction: A New Paradigm of Research in Robotics. *Advanced Robotics*, Vol. 8, No. 4, pp. 357–369, August 1994.

[3] H. Mannila, H. Toivonen, and A.I. Verkamo. Discovering frequent episodes in sequences. In *KDD-95 Proceedings. First International Conference on Knowledge Discovery and Data Mining*, pp. 210–215, 1995.

[4] Beckmann, Kriegel N., R H.P., Schneider, and B. Seeger. R*-tree: An Efficient and Robust Access Method for Points and Rectangles. In *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, pp. 322–331, 1990.

[5] N Katayama and S Satoh. The SR-tree: An Index Structure for high-Dimensional Nearest Neighbor Queries. In *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, pp. 369–380, 1997.

[6] Roger Weber, Hans-Jrg Schek, and Stephen Blott. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *Proc. of Very Large Databases*, pp. 194–205, 1998.

[7] Eamonn J. Keogh, Kaushik Chakrabarti, Sharad Mehrotra, and Michael J. Pazzani. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. In *Proc. of ACM SIGMOD Int. Conf. on Management of Data*, 2001.

[8] http://research.nii.ac.jp/~katayama/homepage/research/srtree/Japanese.html.

[9] http://www.physionet.org/.