

データストリームのためのパターン検出

背景・課題：大量の時系列データストリーム(Webアクセスログやセンサデータなど)からトレンド検出や分類、異常検出などのために共通するパターンをリアルタイムに検出する手法が求められています。しかし、従来の手法では計算コストがデータ長に比例して増大するため、リアルタイム検出が困難でした。

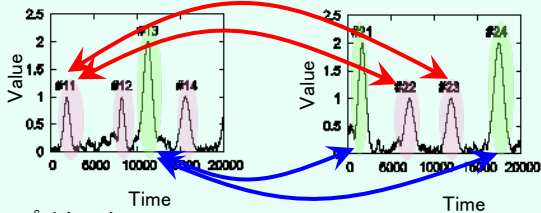
アプローチ：部分シーケンスパターンの間の類似性を評価するために、従来ダイナミックタイムワーピング距離で行っていた類似判定をオリジナルのスコアリング関数に基づく類似判定に変更。類似度が高いほどスコアが上昇するという仕組みにより、似ていないパターンを効率的に排除することができました。

到達点：検出精度を犠牲にすることなく、計算時間とメモリ使用量を大幅に削減することに成功。データを蓄積することなく、到着と同時に一定の計算コストで処理を行います。蓄積されたデータセットに対しても、数千行のWebアクセスログ間の共通パターンを1秒未満で検出できるようになりました。

概要：

研究目的

データストリームからの類似部分シーケンスの検出
・事前の知識なくデータが到着する度にリアルタイムに処理

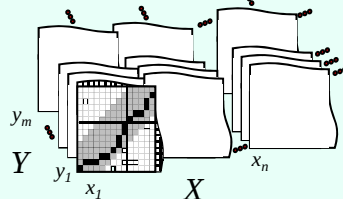


Value
Time

アプリケーション
・Webクリックストリームにおける類似アクセスパターン検出
・センサネットワークにおける類似変動検出

従来手法

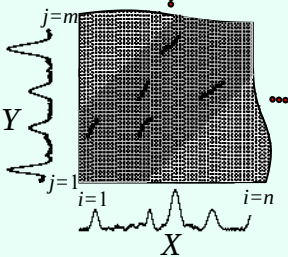
ダイナミックタイムワーピング距離を利用



多くの類似度計算行列が必要
・計算コスト&メモリ使用量：線形増加 $O(nw^2 + mw^2)$

本手法

オリジナルのスコアリング関数を利用

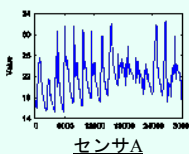


1枚の類似度計算行列のみ

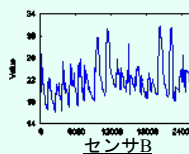
・計算コスト&メモリ使用量：一定値 $O(w)$

※ m, n ：ストリーム X と Y の長さ
 w ：ワーピングスコープの幅

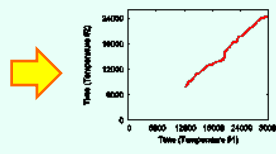
パターン検出例



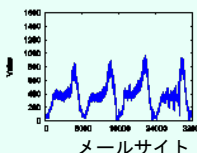
センサA
温度センサから取得されたデータ



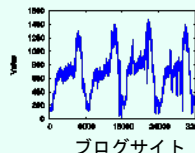
センサB



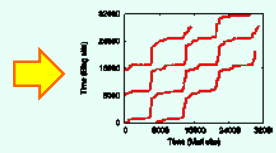
センサ間に隠れた類似する温度変化を発見



メールサイト
Webサイトのアクセス数を記録したデータ



ブログサイト



メールサイトとブログサイトのアクセスパターンの類似性を事前の知識なく検出

関連文献

Machiko Toyoda, Yasushi Sakurai, "Discovery of Cross-Similarity in Data Streams," Proc. IEEE Int. Conf. on Data Engineering (ICDE 2010) pp. 101-104, 2010.

豊田真智子, 櫻井保志, 市川俊一, "データストリームのための部分シーケンスマッチング", DEIMフォーラム2009, E8-7, 2010. (最優秀論文賞受賞)

連絡先

豊田 真智子 (Machiko Toyoda), 櫻井 保志 (Yasushi Sakurai)

協創情報研究部 知能創発環境研究グループ



櫻井

豊田

