数学と情報科学で解き明かす多様な対象の数理構造と活用
2019 年度採択研究者

2019 年度
実績報告書

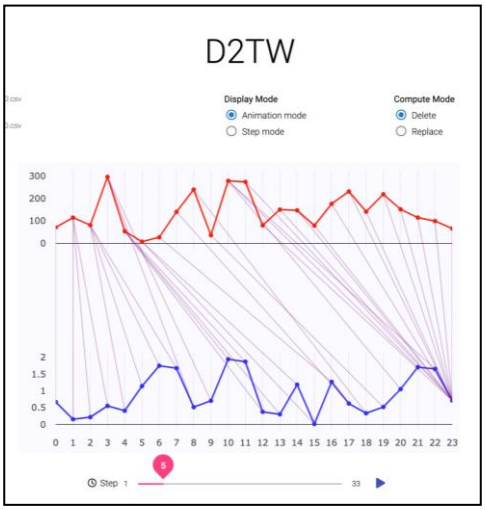
稲永　俊介

九州大学大学院システム情報科学研究院
准教授

文字列学的手法によるシーケンシャルデータ解析

§1. 研究成果の概要

シーケンシャルデータとは，何らかの規則あるいは意味を伴った順序を持つ単位データの列の総称であり，自然言語テキスト，DNA 配列，ウェブ／SNS，センサ観測データ，サーバ通信ログ，ストリーミング配信データ，時系列データなどがその例に挙げられる．本研究では，多様なシーケンシャルデータを，記号列すなわち文字列とみなし，文字列組合せ論とアルゴリズム・データ構造技術の融合によって，統一的かつ高速に処理する基盤技術を開発する．



今年度は，まず，動的時間伸縮距離（DTW 距離）をデータの編集操作に応じて高速計算する手法を開発した．DTW 距離は，データマイニング，ロボット工学，音楽情報処理などの多分野で応用されており，本技術の適用可能範囲は広い．また，現在，本技術をコアとしたソフトウェアを開発中である（図参照）．加えて，DTW と並んで広く用いられる文字列比較指標である LCS の発展形として，解に関するユーザの前提知識を反映可能な STR-EC-LCS を高速計算するアルゴリズムを考案した．

データが逐次送られてくるストリーミングデータに対しては，スライド窓による処理が効果的かつ実用的である．本研究では，入力列上のスライド窓に対する極大ユニークパターン（MUS），および極小不在パターン（MAW）の数理的性質に関する基礎研究を行った．これらのパターンのスライド窓における出現数の厳密な上下界を与えたのち，MUS を最適時間・領域で計算するアルゴリズムを与えた．

上述したデータの編集操作やスライド窓に対する処理を効率行うに際して，動的なデータに対する索引構造が極めて有用である．本研究では，理論・実用の両面において，現在世界最高速かつ

最省領域なコンパクト木索引構造 c-trie++ の開発に成功した.