学位報告

報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　　データストリームにおける
高速パターン検出技術の研究

氏　　名　　豊田　真智子

論　文　内　容　の　要　旨

ネットワークから流入する大量のデータであるデータストリームは，金融データ分析，センサネットワークモニタリング，移動体追跡，Web クリックストリーム分析，ネットワークトラフィック分析といった様々な分野で発生する．このようなデータに対する問合せやマイニングはとても重要なタスクである．ストリームマイニングは途切れることなく到着するデータストリームにおいて，モデルやパターンを表す知識構造を抽出することを目的としている．これは，従来のデータマイニングのようにデータを一度ディスクに格納し，一括して分析するのではなく，リアルタイムに到着するデータストリームを，メモリ上でオンラインに分析しなければならない．そのため，高速な処理が要求されるだけでなく，省メモリで動作するアルゴリズムを考案する必要がある．

部分シーケンスマッチングはデータストリームを継続的にモニタリングする技術の 1 つであり，与えられた問合せシーケンスに類似する部分シーケンスをデータストリームから検出する．この問題は，検出したいパターンが予め決まっている場合に対応できる．一方，部分シーケンスマッチングのもう一つの課題として，両方のシーケンスがデータストリームである時，これらに共通するパターンをどのように特定するかという問題がある．これは，トレンド検出やクラスタリング，異常検出などにも応用的に運用できる．シーケンスマッチングでは，ユークリッド距離を代表として様々な類似尺度が利用されてきたが，ストリームのサンプリングレートの違いや周期の変化に対応する必要があるため，これらに適した類似尺度が望まれる．ダイナミックタイムワーピング(DTW: Dynamic Time Warping)距離は，時間軸方向にシーケンスを伸縮して適合させるため，データストリームにおけるシーケンスマッチングに適した類似尺度である．しかし，DTW は蓄積されたデータセットに対してさえ計算コストが高いことで知られており，データストリーム処理を必要とする本問題に適用することは困難である．

本論文では，データストリームにおいて類似する共通パターンを検出する CrossMatch を考案した．CrossMatch は，計算コストを削減するため，DTW 距離を間接的に計算するスコアリング関数，データストリーム処理の中で，類似する共通パターンの位置を特定する位置行列，類似

する共通パターンを検出すると同時にユーザに提示するストリームアルゴリズムの 3 つのアイディアで構成される．理論的に分析することで，CrossMatch が精度を犠牲にすることなく，少ないリソースで動作することを示す．また，実データと人工データを用いた実験により，CrossMatch の有効性を検証する．

一方，CrossMatch で使用する距離関数を DTW から編集距離に置き換えることで，記号シーケンスにおけるシーケンスマッチングにも適用可能である．この特徴を利用して，CrossMatch をセキュリティ監査におけるパターン検出問題に応用する．近年，企業や自治体などの内部犯行による組織情報の漏えいが社会問題となっており，その抑止対策としてユーザの端末操作ログを定期的にセキュリティ監査する方法が注目されている．現状のセキュリティ監査は，一定期間のログをサンプリングにより選択し，それらを調査する方法で実施されており，その多くが監査人のスキルに依存している．監査ログ量の増加により監査人への負担が増大するため，効率的に監査する手法が求められる．そこで，監査人を支援するため，CrossMatch をセキュリティ監査のために改良した危険行動検出方式を考案した．本方式は，危険行動に該当するパターンを端末操作ログから高速に検出し，監査対象ログの全件探索を達成する．本方式を自治体職員の端末操作ログに適用し，効率的に危険行動パターンを検出できることを確認する．