様　式　Ｆ－１９－２

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ７ 日現在

機関番号：３２６８２
研究種目：国際共同研究加速基金（国際共同研究強化）
研究期間：2017～2022
課題番号：１６ＫＫ０００８

研究課題名（和文）大規模データストリーム解析における入力データとプログラム挙動のモデル化（国際共同研究強化）

研究課題名（英文）Characterization of data and application behaviors of large stream data analyses (Fostering Joint International Research)

研究代表者
秋岡　明香（Akioka, Sayaka）

明治大学・総合数理学部・専任教授

研究者番号：９０３３３５３３

交付決定額（研究期間全体）：（直接経費）　　9,100,000円
渡航期間：　　５ヶ月

研究成果の概要（和文）：本研究の問題意識は、科研費の他の大規模データストリーム解析のプロジェクトに取り組む中で、入力データの特徴により解析プログラムの挙動が大きく変わることが分かったことから生まれた。こうした傾向は他のアプリケーションでも見られる。特に疎行列計算では入力行列の特徴によって行列の前処理手法を変えるといったことが日常的に行われている。そこで、疎行列計算の専門家と問題を共有し共同研究を進めた。また、コロナ禍で共同研究の続行が難しい状況においては、データストリーム解析以外のアプリケーションも様々に検討し、その特徴差分からモデル化するアプローチも試みた。しかし研究期間内に明確なモデルを得ることはできなかった。

研究成果の学術的意義や社会的意義
昨今、社会的重要性を高めている機械学習や強化学習のプログラム的特徴は、本プロジェクトで対象とした大規模データストリーム解析とよく似ている。機械学習や強化学習は、その学習過程で並列化による高速化を行うことが難しい部分も多く、部分的な高速化しか成し得ていない。また、効率よく優れたモデルを獲得するためには、学習データの順序や選択について、知見に基づいた試行錯誤が必要な場合も多い。つまり、大規模データストリーム解析と入力データの挙動をモデリングすることは、機械学習や強化学習の高速化や効率化に繋がる。チャレンジングな問題ではあるが、様々なアプローチを模索しながら、引き続きこの問題に取り組んで行きたい。

研究成果の概要（英文）：The problem of this project arose from the experience in another large-scale data stream analysis project. That is, the behavior of analysis programs can vary significantly depending on the characteristics of the input data. Such tendencies can also be observed in other applications. In particular, in sparse matrix computations, it is quite common to switch the preconditioners based on the characteristics of the input matrix. Therefore, the collaboration with experts in sparse matrix computations is expected to be a great way to attack the project target. In the COVID-19 pandemic, the collaboration became difficult. Therefore, we explored various applications beyond data stream analysis, and attempted to model the target based on their characteristic differences. However, we were unable to obtain a clear model by the end of this project.

研究分野：並列分散処理

キーワード：大規模データストリーム解析　モデル化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

監視カメラ映像や SNS の投稿など，時系列に沿って到着するデータ列をリアルタイム解析するストリーム解析処理を高速化・スケールアウトする計算機環境への要望が高まっているが，その挙動は明確でない．ハイパフォーマンスコンピューティング分野（HPC）では，膨大なデータを解析する処理を研究してきたが，HPC の大規模データ解析処理とストリーム解析処理ではデータアクセスパターンが全く異なるため，高速化の手法は根本的に異なるはずである．HPC での大規模データ解析処理は，利用する計算機環境や利用者が固定的である場合が多い．一方で大規模ストリーム解析処理では，ひとつの解析手法について様々な実装が存在し，利用者は経験に基づいて目的に合った実装を選択する．大規模ストリーム解析処理の高速化・スケールアウトでは，不特定多数の利用者や計算機環境を前提として議論する必要がある．現在の計算機環境は，CPU 性能に強スケールし，処理の主要部分でデータの再利用が頻発するベンチマークを速くすることを大きな目標として発展してきた．しかし，ストリーム解析処理のようにデータの再利用がほとんどなく，CPU 性能に強スケールしづらいアプリケーションにとって，現計算機環境は必ずしも好都合ではない．

大規模データ解析向けの既存ベンチマークを網羅的にサーベイしたが，1)入力データの特徴とプログラムの挙動の関係性を明らかにしたベンチマークは皆無であった，2)ストリーム解析処理のモデル化はほぼ未着手であった，という問題が明らかになった．

２．研究の目的

大規模行列計算では，対象とする行列の特徴に応じて，適用するアルゴリズムを変更したり，前処理を行なったりすることで，計算精度や処理速度の向上を計る必要があり，そのノウハウは長年研究されている．入力データの特徴がプログラムの挙動やアルゴリズム選択に大きな影響を与える点で，大規模数値計算とストリーム解析処理は共通しており，問題解決への根本的なアプローチには共通点があると考える．そこで，大規模数値計算分野の知見を導入し，ストリーム解析処理の入力データの特徴抽出とモデル化を目指した．

３．研究の方法

本研究に着手した当初は，大規模ストリーム解析プログラムと入力データの相関を直接解き明かそうとして，困難を極めた．したがって，初年度末にはアプローチを少し変更し，大規模ストリーム解析に限らない様々なアプリケーションの挙動と入力データの相関を検討することで，その成果を大規模ストリーム解析にフィードバックし，モデルを獲得することを試みた．具体的には，当初から計画していた疎行列計算のほか，流体計算，生物学における分子構造と機能の予測，人体の内臓パーツや関節をシミュレーションするためのメッシュ生成，日本語解析，物体認識，HPC 分野で長年重要視されている SPEC ベンチマークなどである．理論面からのヒントを得るために，応用数学の専門家とディスカッションする機会も何度か持った．

４．研究成果

本研究の問題意識は，科研費の他の大規模データストリーム解析のプロジェクトに取り組む中で，入力データの特徴により解析プログラムの挙動が大きく変わることが分かったことから生まれた．こうした傾向は他のアプリケーションでも見られる．特に疎行列計算では入力行列の特徴によって行列の前処理手法を変えるといったことが日常的に行われている．そこで，疎行列計算の専門家と問題を共有し共同研究を進めた．また、コロナ禍で共同研究の続行が難しい状況においては，データストリーム解析以外のアプリケーションも様々に検討し，その特徴差分からモデル化するアプローチも試みた．こうしたアプローチの変更は，研究期間内に議論した他の情報科学研究者から，問題の重要性について賛同を得ると同時に，問題解決が非常に難しいという指摘も受けたことを反映した結果であった．しかし研究期間内に明確なモデルを得ることはできなかった．

本研究計画が開始したのは 2017 年であるが，そこから 5 年以上の月日が経過し，機械学習を取り巻く状況も大きく様変わりした。トラディショナルなシミュレーション等による予測よりも，機械学習を用いた手法の方が圧倒的に高速に許容解に辿り着く傾向が大きくなった．また，機械学習の学習過程においても，大量のデータで様々な学習ルートを実施し，精度が高いモデルを作成して公開する動きも増えた．こうした精度の高いモデルは転移学習等にも用いられ，多くのユーザは１から学習モデルを構築する必要がなく，既存のモデルの利用や転用で，十分な目的を果たすことができるようになった．本来であれば、こうした多くの利用者が付いている学習済みモデルの構築過程で得られたであろう知見を集めて，学習のモデル化を行うことができれば良かったのかもしれないが，基本的に大量のデータで押し切ることが前提となっている傾向もあり，

世の中のこうした流れを自身の研究に取り入れることは難しかった.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　0件）

１．発表者名
吉澤康太，秋岡明香

２．発表標題
マイクロブログ情報解析によるイベント抽出手法の提案

３．学会等名
第82回情報処理学会全国大会

４．発表年
2020年

１．発表者名
松田莉奈，秋岡明香

２．発表標題
構図情報に注目した画像クラスタリング

３．学会等名
第82回情報処理学会全国大会

４．発表年
2020年

１．発表者名
松永有紀子，秋岡明香

２．発表標題
刀の形状推定

３．学会等名
第82回情報処理学会全国大会

４．発表年
2020年

１．発表者名
小林可奈子，秋岡明香

２．発表標題
キータイピングに表れるユーザ個人の特徴を利用した文書改ざん防止エディタ

３．学会等名
第82回情報処理学会全国大会

４．発表年
2020年

１．発表者名 秋岡明香
２．発表標題 意味ベクトル表現を用いたJ-POP歌詞の文体分析
３．学会等名 第82回情報処理学会全国大会
４．発表年 2020年

１．発表者名 大貫佑真，秋岡明香
２．発表標題 画像処理を用いた点字学習支援アプリケーション
３．学会等名 第82回情報処理学会全国大会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
主たる渡航先の主たる海外共同研究者	ションツ　スザンヌ (Schontz Suzanne)	カンザス大学・Dept. of Electrical Engineering and Computer Science・Professor	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
米国	University of Kansas			
米国	University of Kansas			
米国	University of Kansas			