様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １１ 日現在

機関番号：３２６８２
研究種目：若手研究(B)
研究期間：2015～2019
課題番号：１５Ｋ２１４２３

研究課題名（和文）大規模データストリーム解析高速化に向けたベンチマークセットの構築

研究課題名（英文）benchmark set to speed up extremely large scale data analysis applications

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）　　3,000,000円

研究成果の概要（和文）：本研究課題では、大規模データストリーム解析を高速化する手がかりとしてのベンチマークセットを構築することを目指した。具体的には、これらのプログラムを挙動に合わせて分類し、各分類ごとに特徴的かつ代表的な挙動をモデル化し、このモデルに沿った動きをする一連のプログラム群を構築・公開することを目指した。結論としては、ベンチマークセットの構築には至らなかった。当初の想定以上に入力データの違いがプログラムの挙動の違いを招き、その種類は極めて多岐にわたること、入力データが持つどのような特徴が挙動に影響を与えるかは、投機的実行なしでは推定が困難であること、データのモデル化が不可欠であることがわかった。

研究成果の学術的意義や社会的意義
大規模データ解析は現在でも重要で注目を浴びているアプリケーションであり、中でもストリーム解析はリアルタイムに大規模なデータ解析を行う上で不可欠である。このように社会的需要が高いアプリケーションを高速化することは、新たな応用分野の開拓や新しい知見を得るために不可欠であるが、抜本的な解決法は発見されていない。本研究でも問題解決に至ることはできなかったが、問題解決に向けて取り組むべき問題を以前より具体的にすることはできた。

研究成果の概要（英文）：This project aims to build a benchmark suite, which is expected to help speeding up data stream analysis, with special focus on the extremely large scale input data. More concretely, categorization of streaming analysis is expected to introduce characteristic behaviors of the programs, and representative examples for each category are expected to form a benchmark suite for this community. After the period of this research project, the benchmark suite could not be completed. Struggles during the research period brought several directions for the further research. For example, input data has much influence than expected, and the variety of the pattern is enormous. Extraction of dominative characteristics in input data is a hard problem, and speculative execution is the current solution. Modeling of input data is unavoidable for modeling of programs in this area.

研究分野：並列分散処理

キーワード：ベンチマーク

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

大量のデータを横断的かつ網羅的に解析し、新しい知見やニーズを発見しようとするビッグデータ解析が注目を集めている。こうした需要や期待の高まりに応じて、ビッグデータ解析処理を高速化・スケールアウトする新たな計算機環境への要望も生じているが、ビッグデータ解析処理の挙動は、未だ明確でない。

ビッグデータ解析処理はその特徴によって分類できるが、本提案はストリーム解析処理に注目する。ストリーム解析処理とは、時系列に沿って次々と到着するデータ列を（ほぼ）リアルタイムに解析する処理を指し、ストリーム解析処理に特化したアルゴリズムはデータマイニング分野で集中的に研究されている。ハイパフォーマンスコンピューティング分野（HPC）においても、膨大なデータを解析する処理は長年研究されてきたが、ストリーム解析処理と HPC の大規模データ解析処理とでは、データのアクセスパターンが全く異なるため、高速化の手法が根本的に異なる。

また、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く、しばしば特定の条件下で高速化・スケールアウトすることが目標となる。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によるライブラリが存在し、利用者はこれらのライブラリから経験に基づいて目的に合った実装を選択して利用する。したがって大規模ストリーム解析処理では、不特定多数の利用者や計算機環境を前提として、高速化やスケールアウトの指針を議論する必要がある。

さらに、現在の計算機環境は Linpack や SPEC などの主に 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のひとつ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、その性能が 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、現在の計算機環境は好都合ではない。

２．研究の目的

本提案では、以下の 4 項目について研究・開発を行なう。

(1) ストリーム解析処理プログラムのデータ依存関係や処理依存関係を解析するツール
(2) ストリーム解析処理の代表的な各手法で頻繁に用いる入力データのデータモデル
(3) ストリーム解析処理の代表的な各手法のデータ依存関係や並列化可能部分を表すデータ依存グラフ
(4) 同様に、処理依存関係や並列化可能部分、計算コストを表す処理依存グラフ

ストリーム解析処理の代表的な手法ごとに、(2)、(3)、および(4)をまとめたものを、その手法のベンチマークセットとして公開する。一般的に、(3)のデータ依存グラフ、および(4)の処理依存グラフの組み合わせをタスクグラフと呼び、HPC においては対象アプリケーションの特徴を示すモデルとして広く利用されている。また、一般的なタスクグラフはデータモデルを含めないが、ストリーム解析処理では同一手法でも入力データによって実行時間が大きく異なることが分かっていることから、本提案ではタスクグラフにデータモデルを組み合わせてベンチマークセットとする。

３．研究の方法

本提案の研究課題は、ストリーム解析処理の代表的手法について、入力データをモデル化し、さらに各入力データモデルを使用した場合のストリーム解析処理手法の挙動をモデル化したタスクグラフを生成することである。そして、頻繁に使用する入力データおよび代表的なストリーム解析処理手法を網羅的に解析し、各入力データモデルおよびタスクグラフの組み合わせが持つ特徴によって分類を行ない、最小かつ十分な数の組み合わせに絞り込むことである。

初年度はタスクグラフ作成自動化のためのツール構築に注力し、大規模かつ効率的に入力データのモデル化とタスクグラフ生成を行なう基盤を固める。2 年目以降で集中的に入力データのモデル化とタスクグラフ生成を進め、3 年目後半から最終年度にかけて、入力データモデルとタスクグラフの組み合わせの絞り込みを行なうことで、ベンチマークセットとしてのタスクグラフセット構築を目指す。

４．研究成果

本研究課題では、大規模データストリーム解析を高速化する手がかりとしてのベンチマークセットを構築することを目指した。具体的には、これらのプログラムを挙動に合わせて分類し、各分類ごとに特徴的かつ代表的な挙動をモデル化し、このモデルに沿った動きをする一連のプロ

グラム群を構築・公開することを目指した。結論としては、ベンチマークセットの構築には至らなかった。当初の想定以上に入力データの違いがプログラムの挙動の違いを招き、その種類は極めて多岐にわたること、入力データが持つどのような特徴が挙動に影響を与えるかは、投機的実行なしでは推定が困難であること、データのモデル化が不可欠であることがわかった。

大規模データ解析は現在でも重要で注目を浴びているアプリケーションであり、中でもストリーム解析はリアルタイムに大規模なデータ解析を行う上で不可欠である。このように社会的需要が高いアプリケーションを高速化することは、新たな応用分野の開拓や新しい知見を得るために不可欠であるが、抜本的な解決法は発見されていない。本研究でも問題解決に至ることはできなかったが、問題解決に向けて取り組むべき問題を以前より具体的にすることはできた。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Sayaka Akioka, Suzanne M. Shontz
２．発表標題 Dataset Characterization of Data Mining Algorithms
３．学会等名 SIAM Parallel Processing (SIAM PP 2018)（国際学会）
４．発表年 2018年

１．発表者名 Sayaka Akioka
２．発表標題 Benchmarking Big Data Applications: A Review
３．学会等名 The Eighth International Conference on Future Computational Technologies and Applications（国際学会）
４．発表年 2016年

１．発表者名 Sayaka Akioka
２．発表標題 Stream Mining Revised
３．学会等名 The Second International Conference on Big Data, Small Data, Linked Data and Open Data（国際学会）
４．発表年 2016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