

研究終了報告書

「超高速な多モーダル IoT データ統合処理基盤」

研究期間:2020 年 11 月～2024 年 3 月

研究者:塩川 浩昭

1. 研究のねらい

デバイスからリアルタイムに生成されるデータはとどまることがなく、そのデータ量も極めて膨大となる。総務省情報通信白書では、ひとりあたりのデバイス数は近い将来 1,000 台を超えることが示唆されており、そのデータ量はさらに増加することが見込まれている。各種デバイスを経由して時々刻々と生み出される IoT ビッグデータを高速・高精度に分析することの重要性については疑いの余地がない。一方で、現在の IoT 環境が生み出す膨大なデータ量と我々が日常的に利用する計算資源の間には大きなギャップが生じている状況にある。従来の IoT データ処理では取得したデータを高性能なサーバに集約し計算することが一般的であるが、多くのデータ分析の実運用現場において高性能計算機を利用できる場面は珍しく、実際はユーザの保有する性能の限定された計算環境でデータを処理する必要がある。ところが、演算能力・メモリ容量といった時間的・空間的制限を持った計算資源上においてビッグデータを扱う方法は自明ではなく、このような環境下における高速ビッグデータ処理の実現が大きな命題となっている。

本研究のねらいは、IoT データに対する超高速なデータ処理アルゴリズム基盤を開発することで、IoT ビッグデータ処理における計算機の時空間的な制約を克服することである。具体的には、リアルタイムかつ多モーダルな IoT ビッグデータに対して、(1)限られた計算資源での超高速処理を実現する新しいアルゴリズム理論の開拓、(2)このアルゴリズム理論に基づく超高速な IoT データ分析手法の開発、ならびに(3)本研究を通じて開発した技術のライブラリ化や公開を行うことで、実際の大規模データ分析アプリケーションやユースケースでの実践・実利用を進め、ユーザからのフィードバックを得る。前述した通り、IoT ビッグデータはその計算量の高さから高性能な計算機を必要とする場合が多い。しかしながら、我々は必ずしも高性能計算機を利用できるわけではなく、両者には大きなギャップが存在する。本研究では新しいアルゴリズム理論を核として、超高速な分析技術を開発していくことでこのギャップを解消し、IoT データ処理の時間的・空間的制約を克服することを目指す。

2. 研究成果

(1)概要

本研究は(1)理論開拓、(2)アルゴリズム開発、(3)実践・展開、および(4)領域内外との連携の4つの観点においてそれぞれ以下の主たる成果を獲得した。

(1) 理論開拓

大規模な IoT データを高速処理するために、グラフモデルを介したデータの頻出構造（以降、基調構造）の抽出とそれを利用したアルゴリズムの効率化メカニズムを理論的に解明した。静的なグラフを対象とした探索問題を題材とし、基調構造に基づくアルゴリズム設計方式は最先端手法と比較して最大 10,000 倍高速に計算可能となることを明らかに

した。また、一定の条件下における時空間的な性能限界を理論的に明らかにした。本成果の一部は難関国際会議 IJCAI 等で発表を行った。

(2) リアルタイムアルゴリズム開発

上述の設計方式に基づき多モーダルデータ処理アルゴリズムの高性能化を実現した。異種データ類似検索や範囲検索、多次元時系列データに対する類似検索やクラスタリング、分散環境における相関時系列探索などを対象とした高速化に取り組み、いずれの対象に対しても最先端の 100 倍～30,000 倍程度の高速化を達成した。本成果の一部は難関国際会議 AAAI, ISWC, IJCAI 等で発表を行った。

(3) 実践・展開

開発したアルゴリズム群の有効性を実運用現場において検証するとともに、OSS ライブラリとして展開した。具体的には IoT を活用した創薬分野において本研究のアルゴリズム群を展開し、これらは従来方式よりも 1,000 倍程度高速かつ高精度に化合物を検索できることを明らかにした。

(4) 領域内外との連携

本研究の進展に伴い、領域内においていくつかの研究連携が実現した。代表的なものは、松井勇佑氏（東大）、天方大地氏（阪大）との連携であり、このプロジェクトは現在 AIP 加速プログラムに採択されている。また、領域外での連携では(i)創薬分野への成果応用に関する連携、(ii)グラフデータベースへの成果応用とその標準化に関する国際研究連携、(iii)複数の企業との研究連携が実現した。

(2) 詳細

研究提案の実施計画は大別して次に示す 3 つのフェーズで構成された。「(1)理論開拓」では本研究の基礎となる基調構造を捉えた多モーダルデータ処理方式の開発し、その性能限界を理論的に解明した。「(2)アルゴリズム開発」では前述した方式に基づいた高性能な（リアルタイム・分散処理）アルゴリズム群の開発を行った。「(3)実践・展開」では、開発したアルゴリズム群の有効性を実運用現場において明らかにするとともに、その成果をライブラリとして公開し、広く展開した。各フェーズの主要な研究成果について以下に述べる。

1. 理論開拓

(成果 1.1) 最大 k-plex 検出を題材とした高速データ処理性能の理論的限界解析

大規模グラフの中から最大の k-plex を高速に検出する手法を開発した。最大 k-plex 検出は NP 困難問題であり、大規模データに対して現実的な計算時間で処理することが難しい。この問題に対して、本研究では実世界のグラフデータが持つ基調構造を捉えることで、データ件数 n に対して従来 $O(n^2) \sim O(n^3)$ の計算時間を要する処理を $O(n)$ 程度に抑制可能であることを理論的に明らかにした。また、この理論的な性質に基づき最大 k-plex の高速検出手法 BnM を開発した。BnM は数千万件規模のデータを対象として、その最大 k-plex を最先端手法よりも最大で 2,300 倍程度高速に検出可能である。図 1 はノード数を変化させた際の計算時間の推移を比較した結果である。提案手法 BnM は他の最先端手法と比較したとき、大規模データに対しても十分に高速に最大 k-plex を検出することが出来ている。本研究の成果は IJCAI2023 に採択され、発表を行った。

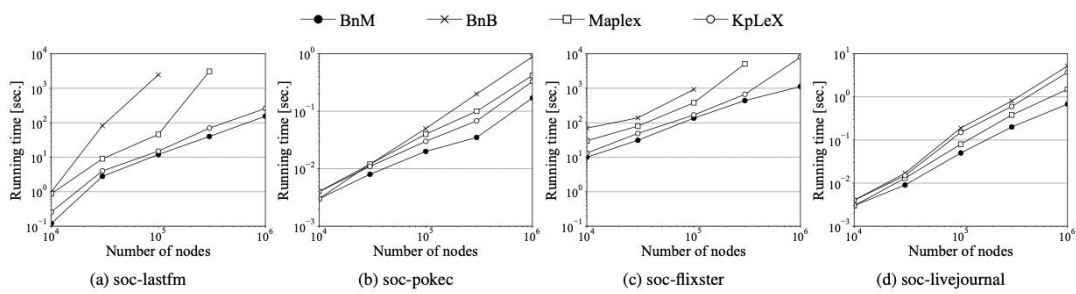


図 1. ノード数の変化に対する最大 k-plex 検出時間の推移

(成果 1.2) 多モーダルデータ処理に対する高速処理法の開発と理論解析

多モーダルデータをハイパーグラフとみなし、そのクラスタリング処理に対して基調構造を用いた高速化手法を開発した。また、基調構造の持つ理論的な性質を解析し、本研究のアプローチの時間・空間計算量上の限界を理論的に検証した。この問題はデータ件数を n 件としたとき、 $O(n^2)$ 時間の計算コスト必要となり、IoT ビッグデータを対象とした際に膨大な計算時間を必要とする。本研究ではこの問題に対して、基調構造に基づく高速化手法を実装するとともに、階層的なデータ集約手法ならびに基調構造の枝刈り手法を導入し、提案方式の有効性を検証した。図 2 はグラフのノード数の変化に対する提案手法(Ours)と最先端手法の実行時間の比較である。提案手法はどのノード数に対しても最先端手法よりも約 10,000 倍以上高速であることを明らかにした。また、図 3 では正解クラスタレブルに対する精度を比較したが、精度についても提案手法が他の手法を上回る結果となった。以上の結果から、多モーダルデータ処理においても、基調構造を用いた高速化方式は 10^7 件規模の効果的に機能することを明らかにした。

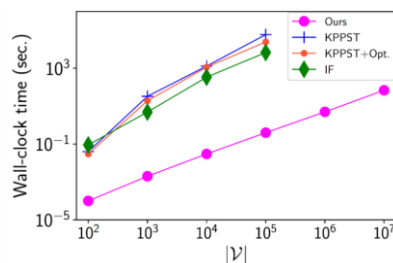


図 2. 実行時間の比較

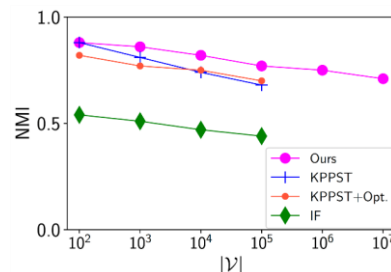


図 3. クラスタリング精度の比較

2. アルゴリズム開発

(成果 2.1) 多モーダル IoT データに対する類似検索の高速化

多モーダルデータをオントロジーによりアノテートし、より高度な学習モデル構築に活用する事例が増加している。しかし、アノテートされた大量の情報に対する検索は大きな計算コストを必要とする。本研究ではスキーマ情報を活用して、このような環境下における類似検索処理の高速化・省メモリ化手法を開発した。提案手法はスキーマ情報から属性間の基調構造を捉え、検索処理コストを削減する。3,660 万件規模のデータにおいて、提案手法は最先端手法より 100 倍～300 倍程度高速に検索できることを確認した。また、提案手法は高速化とともに従来手法よりも高い検索精度達成することを明らかにした。本研究の成果は ISWC2021 に採択され、発表を行った。また、提案手法は OSS ライブラリとして GitHub 等を通じて公開した。

(成果 2.2) 多モーダル IoT データに対する範囲検索の高速化

本研究では多モーダルな属性を持った IoT データを統合管理するネットワークに対して、効率的な範囲検索を実現するための索引構築手法を提案した。提案手法では、多モーダルな属性間の基調構造を抽出し、それぞれの基調構造に対して属性索引と基調構造索引を構築する。属性索引と基調構造索引は、それぞれ複雑ネットワークの構造を考慮することで検索に必要な情報を $O(1)$ 時間で見つけることができる。図 4 は提案手法と既存手法の検索時間の比較を行ったものである。図中の CT+BAG は提案手法、CT+Naive および CT+G-Tree は最先端手法を示している。この結果からも明らかなように、提案手法は既存手法と比較して 100～1,000 倍高速な検索処理を実現している。また図 5 に示すとおり、BAG-index は基調構造を捉えることで、索引構築時間も既存手法と比べて 20,000 倍程度高速である。本成果は国際学術雑誌 *Social Network Analysis and Mining* (Springer) にて出版した。

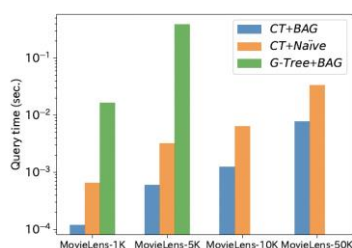


図 4. 問合せ処理時間の比較

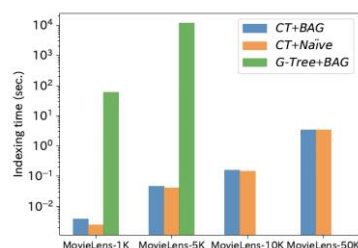


図 5. 索引構築時間の比較

(成果 2.3) 多モーダルオブジェクトクラスタリングの高速化

Affinity Propagation は多次元データを含む多様なデータオブジェクトを対象としたクラスタリングアルゴリズムである。Affinity Propagation は入力データ間で完全グラフを構築し、そのグラフのエッジに対してメッセージパッシングを行うことで高精度にクラスタを検出する手法である。しかしながら、メッセージパッシング処理はデータ件数を n としたとき、 $O(n^3)$ 時間のコストを必要とする。本研究では基調構造とその決定性を利用することで、必要最低限のデータのみ計算するようにアルゴリズムを開発した。本研究の提案手法は、数百万件程度のデータセットに対して従来手法と比較して精度を劣化させず 100 倍以上高速な処理を可能とした。これに対して既存手法は提案手法よりも低速であり、多くの場合、近似解のみを出力する。この研究成果は AAAI2021 に採択され、発表を行った。また、本研究で開発したアルゴリズムは OSS ライブラリとして GitHub 等を通じて公開した。

(成果 2.4) 多モーダル時系列データに対する類似モチーフ検出

多モーダルな時系列データに対する高性能な類似モチーフ検出手法を開発した。この研究では図 6(左図)のようなクエリモチーフが与えられたとき、クエリモチーフと最も類似した部分多次元時系列データを図 6(右図)に示すような多次元時系列データから高速に列挙する問題を考える。既存の類似モチーフ検出処理では全ての部分系列を検索する処理が必要であり、膨大な計算コストを要する。本研究では時系列データをグラフデータに変換して抽出した基調構造をもとに検索処理の高速化を行った。図 7 に示すように、多モーダルな時系列データにおいて、提案手法は概ね 10 倍程度の高速化を達成していることを本研究では確認した。本研究の成果は iiWAS2023 にて発表を行った。

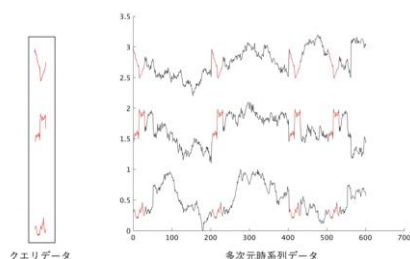


図 6. 類似モチーフ検出例

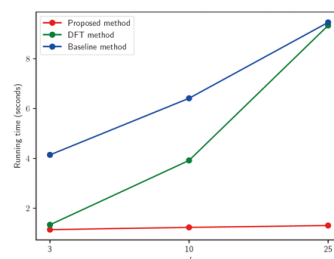


図 7. モーデル数に対する処理時間の比較

(成果 2.5) 高速・省メモリ・高精度な分散データ処理フレームワーク

ネットワーク上に分散配置された計算資源を効率的に活用してリアルタイムデータを高速・高精度・省メモリに処理するためのフレームワークを開発した。本研究では IoT エッジデバイスのような少ないメモリ容量の上で高速・高精度な計算を実現するため、提案フレームワークでは基調構造を利用したデータ圧縮を行う。提案フレームワークはこのビューを分散配置された計算資源に分散配置してクエリに対するデータ処理(検索, 分類, クラスタリングなど)を実行する。各ビューはデータ圧縮されているため、元のデータセットを計算した場合と比較して、高速・省メモリに応答することができる。各ビューの計算結果を集計し、元のデータに対する計算結果を予測する。図 8 では実際のデータセットを用いた実行時間の比較結果を示しているが、提案フレームワーク Corgi は他の方式と比べて最大 90 倍程度高速に処理できることを確認した。本研究の成果は IJCAI2024 に採択された。

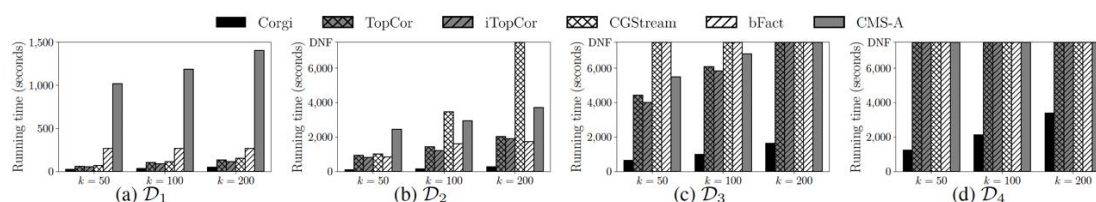


図 8. 計算時間の比較 (DNF は 6 時間以内に処理が完了しなかったことを示している)

3. 実践・展開

(成果 3.1) 創薬分野における実践的応用

創薬・素材探索領域において本研究で開発したアルゴリズムの実応用を実施し、その有効性を検証した。具体的には、創薬の際に非常に重要な処理となる Ligand-based Virtual Screening (LBVS)での開発アルゴリズム群の応用を行った。本研究では LBVS に対して本研究のアルゴリズム群を展開し、従来方式よりも 1,000 倍程度高速に化合物を検索できることを明らかにした。また、アルゴリズムの実用上の有用性を検証するための検索精度評価を行っており、こちらでも極めて良好な結果を獲得した。本研究の成果の一部は iiWAS2023 や IJCAI2024 に採択された。また、本研究で開発した高速・高精度な類似検索アルゴリズムは、化合物データ処理ライブラリとしてのリリースを予定している。

3. 今後の展開

本研究を通じて、多モーダルなデータ処理に対する超高速なデータ処理のための理論やアルゴリズム設計方式については整備されつつある状況にあると考える。本研究の根幹にある基調構造

を捉えた高速化アプローチは、グラフデータや時系列データといった多様なデータ形式に対しても同程度に効果的であることを本研究は明らかにしてきた。この成果を受けた今後の展開としては、より具体的な応用・ユースケース・アプリケーションに特化したアルゴリズムの開拓・整備を進めていくことが目標の一つである。具体的には、上述の「(4)領域内外との連携」でも述べたように、本研究の成果は電波工学、創薬、データベースといった多様な応用領域に展開されつつある状況にある。とりわけ、創薬やデータベースへの展開では、成果のライブラリへの組み込みや標準化といった具体的な普及戦略が進行しており、社会実装に繋がりとつつある状況にある。したがって今後の計画としては、比較的開発サイクルの短い創薬分野への展開では、3～5年程度で応用領域に特化した研究レベルの技術開発を進めつつ、並行してライブラリへの展開を進める予定である。また、標準化活動に関して長い時間を要するデータベース分野への展開では5年程度で研究レベルでの技術開発を進めつつ、5～8年後に標準化提案が採用されることを目指して進めていく計画である。

4. 自己評価

本研究が大きな目標として掲げていた(1)高速化のための基礎理論の開拓、(2)同理論に基づく具体的な高速アルゴリズムの開発、および(3)開発したアルゴリズムを用いた実応用実践と研究成果のライブラリ化については、本研究を通じて十二分に達成することが出来たと考える。本研究で得られた成果の多くは、最難関国際会議・学術雑誌論文(AAAI, IJCAI, ISWC, Social Network Analysis and Mining 等)で発表・掲載され、各研究分野の最もプレゼンスの高いコミュニティにおいて本研究の核となるアイデアやその成果が十分に評価された。また、主要な研究成果については全てオープンソースソフトウェアとして GitHub にて公開し、MIT License のもとで自由に利用可能な状態に整備されており、普及・展開に向けた当初の目標は達成できたと認識している。加えて、実応用実践では創薬分野において顕著な成果を獲得し、当初の予定を超えて、化薬分野への論文投稿や、化合物データ解析ライブラリとしてのリリースを見込む状況となった。このように、社会実装の観点においても本研究は大きな進展があったと評価している。

研究の進め方については、新型コロナウイルスの流行による海外渡航制限が長期化した影響から、一部の国際研究連携については本研究期間の後半へと後ろ倒しせざるを得ない状況になった。しかし、国際研究連携を除いた研究進捗については概ね計画通り進行することができた。また、海外渡航制限が解除された2022年度以降は、海外の研究者のもとへ直接訪問する機会を意識的に増やすよう対応をしたため、上述の成果「(4)領域内外との連携」でも述べたように、長期的な連携が見込める国際的な研究プロジェクトの立ち上げに繋げることができた。

社会・経済への波及効果に関しては、本研究は計画段階において開発したアルゴリズムを用いた実応用実践を通じてその端緒を模索していくことを狙いとしていた。その観点において、本研究の現時点の成果は将来的な社会・経済への一定の波及効果が見込める段階にあると考えている。具体的な理由は2点ある。まず1つ目の理由は、上述の(成果3.1)で示した創薬分野への実応用が進んでいる点にある。研究成果として述べたように、本研究の成果は創薬分野において一定の有効性が確認され、ライブラリとしてのリリースも見込まれている。次に2つ目の理由は本研究の成果を核としたデータベース分野での標準化提案を進めている点である。こちらについては長い時間を要するが、標準化提案の一部として採用された暁には、産業分野への波及効果ももたらし得る成果となると考えている。

5. 主な研究成果リスト

(1) 代表的な論文(原著論文)発表

研究期間累積件数: 24件

1. Shohei Matsugu, Yasuhiro Fujiwara, Hiroaki Shiokawa. Uncovering the Largest Community in Social Networks at Scale. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023). 2023, pp.2251-2260.

大規模なグラフデータの中から k -plex と呼ばれる最大密部分グラフを高速に検出するためのアルゴリズム BnM を提案した論文である。グラフデータの基調構造を捉えることにより, BnM は従来よりも最大 2,300 倍高速な計算を可能とした。また, BnM は既存手法と比較してもデータ処理精度を劣化させることなく高速計算が可能であり, これは本研究の核となるアイデアの有効性を示している。

2. Suomi Kobayashi, Shohei Matsugu, Hiroaki Shiokawa. Indexing Complex Networks for Fast Attributed kNN Queries. Social Network Analysis and Mining. 2022, Vol.12, No.82, 21 pages.

多モーダルな属性を持った複雑ネットワークに対して, 効率的な k 最近傍問合せを実現するための索引構築手法を提案した。提案手法 BAG-index では, 複雑ネットワークに含まれる基調構造を抽出し, それぞれの基調構造に対して属性索引と基調構造索引を構築する。提案手法は既存手法と比較して 100~1,000 倍高速な検索処理を実現した。

3. Hiroaki Shiokawa. Fast ObjectRank for Large Knowledge Databases. In Proceedings of the 20th International Semantic Web Conference (ISWC 2021). 2021, pp.217-234.

本論文では IoT デバイスから取得された多モーダルなデータを統合処理するためのデータモデルであるヘテロジニアスグラフを対象とした高速なデータ検索アルゴリズムを開発した。モーダルの異なる複数のデータを高速・省メモリに統合処理するために, 提案アルゴリズムではデータ間の関係性を表したスキーマグラフを構築して基調構造を捉える。これにより, 従来よりも 100 倍以上高速な検索処理を実現した。

(2) 特許出願

研究期間全出願件数: 2 件(特許公開前のものは件数にのみ含む)

1	発 明 者	塩川 浩昭, 真次 彰平, 藤原 靖宏
	発 明 の 名 称	探索装置, 探索方法及びプログラム
	出 願 人	国立大学法人筑波大学, 日本電信電話株式会社
	出 願 日	2023 年 5 月 31 日
	出 願 番 号	特願 2023-089910
	概 要	本発明は, 与えられたグラフにおける最大の「 k -plex」を探索する速度を向上させることが可能である探索装置, 探索方法及びプログラムを提供することを目的としている。

(3) その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

1. Hiroaki Shiokawa, “Scalable Affinity Propagation for Massive Datasets,” In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), 35(11), pp.9639-9646, 2021.

2. Hiroaki Shiokawa, “Massive Graph Analysis at Scale,” The 18th Korea-Japan Database Workshop (KJDB 2023), October 27th-29th, 2023. (招待講演)
3. Hiroaki Shiokawa, “Fast Similarity Search for Large Knowledge Graphs,” Aarhus Workshop on Graph Access and Analysis (AGAN Workshop), June 29th, 2023 (招待講演)
4. Hiroaki Shiokawa, “Graph-based Clustering at Scale,” Aarhus Workshop on Graph Access and Analysis (AGAN Workshop), June 27th, 2023 (招待講演)
5. 第 14 回データ工学と情報マネジメントに関するフォーラム(DEIM2022) 優秀論文賞