公開

（事後評価・研究者用）

研　究　報　告　書

「Data Skewness を捉えた超高速・省メモリな大規模データ処理」

研究期間： 2019 年 4 月～2021 年 3 月
研究者番号： 50225
研 究 者： 塩川　浩昭

1．　研究のねらい

近年，ビジネスや医療，スポーツなどの幅広い分野においてデータ分析技術の活用が成功を収めており，時々刻々と生み出される大規模なデータを高速高精度に分析処理することの重要性については疑いの余地がない．一方で，高速に高い精度の分析処理結果を獲得するためには，高い計算性能を持った計算機が不可欠である．大規模なデータに対して高速かつ高精度なデータ処理を行おうとした場合，それに見合った計算環境を利用者は準備する必要があり，誰でも容易に大規模データを扱うことが出来るわけではないのが現状である．

本研究では，大規模データ処理が必要とする計算資源と我々が手にすることができる計算資源のギャップを埋めるべく，多様な計算環境における高精度なデータ処理を想定した超高速・省メモリな大規模データ処理アルゴリズムの開発に取組む．一般的に我々が容易に入手可能な計算環境は，前述の高性能計算機と比較して CPU やメモリ，バス速度などの性能が低い場合が多い．とりわけ，CPU 性能とメモリサイズは，データ処理の規模と性能に直結する重要な要素であり，両者の性能低下は扱えるデータと分析の規模を直接的に制限する要因となり得る．そこで本研究では，超高速かつ省メモリな大規模データ処理アルゴリズムを開発・提供することにより，誰しもが手持ちの計算環境でビッグデータ処理を実現できるようにすることを目指す．

本研究では実世界のデータの中に含まれているデータ分布の偏りや属性間の従属性などといったデータの偏り（Data Skewness）に着目する．例えば，現実世界に存在するグラフデータには特定の部分グラフ構造が頻出するということがこれまでの研究で明らかになっている．本研究は実データの持つ Data Skewness を捉えることで既存のデータ処理アルゴリズムを再設計し，高速かつ省メモリなアルゴリズム群の構築を目指す．これまでの ACT-I 期間では決定的アルゴリズムの性質に基づいた Data Skewness Caching と呼ばれる高速化手法を提案した．加速フェーズ期間ではこの性質を前提に，(1) 多様なグラフデータ処理，ならびに(2) 全点対計算・データベース処理のような多次元データ処理に対するアルゴリズムの大幅な高速化を狙う．

2．　研究成果

（1）概要

大規模データに対する Data Skewness を捉えた超高速・省メモリなアルゴリズムの構築を目標として，本研究期間では【(1) 属性付きグラフや不確実グラフといった多様なグラフデータに対する分析アルゴリズムの高速化】，ならびに【(2) 全点対計算やデータベース検索処理などを含む多次元データ処理アルゴリズムの高速化】を対象とした研究開発を実施した．これらの研究を通じ，複雑な構造を持つグラフデータや構造の定まらない多次元データに対する処理についても，Data Skewness Caching が有効に働くことを確認することができた．

（2）詳細

【研究課題（1)】多様なグラフデータに対する分析アルゴリズムの高速化

本研究の目的は多様なグラフデータに対して Data Skewness を利用した高速化手法を構築することである．本研究期間では不確実グラフに対する信頼性問合せや属性付きグラフに対するコミュニティ検索を題材として，Data Skewness Caching やデータ構造の偏りを利用した高速化アプローチの開発を行い，本研究の提案する方式の有効性を確認した．数百万ノード規模の実データセットを利用した評価実験を通じて，各題材に対する提案手法が最先端の手法よりも数十～数百倍程度高速に計算可能であることを確認した．また，本研究では上述した手法を応用して分散計算環境におけるグラフデータ処理の効率化にも取り組み，スケーラビリティの高い分散グラフ処理を実現した．以下に各成果の詳細をまとめる．

- 不確実グラフに対する信頼性問合せの高速化

不確実グラフはグラフのエッジに対してその生成確率が付与されたグラフ構造であり，実世界の現象をモデル化するために用いられる重要なデータ構造のひとつである．本研究ではこの不確実グラフにおいて，任意の 2 ノードが接続される確率（信頼性）を高速に計算する手法 Sharing RCSS+ (Sharing Recursive Cut-Set Sampling+) を開発した．Sharing RCSS+は不確実グラフから生成され得るグラフインスタンスをサンプリングすることで信頼性を推定する．この際に，不確実グラフ内で頻出する部分グラフ構造を考慮して不確実グラフを分割することで，サンプリングの対象とする確率空間を分割する．これにより，少ないサンプル数を用いて高精度に信頼性を推定することを可能とする．

提案手法 Sharing RCSS+と既存手法である MC [Fishman, 1985], BFSS [Zhu et al., 2015] とのサンプル数および信頼性推定時間の比較を図 1 に示す．図 1 (a)は推定した信頼性が一定の推定精度に達するまでに必要としたサンプル数の比較である．提案手法はいずれの設定においても既存手法よりも少ないサンプル数で信頼性を推定できていることが確認できる．図 1 (b)は一定の精度水準に達するまでの推定時間の比較結果を示しており，提案手法は推定時間を数十倍～数百倍程度高速化することに成功していることがわかる．この結果からも示唆される通り，提案手法は少ないサンプル数で高精度の信頼性推定を可能とする．本研究ではこの性質を理論的にも証明した．

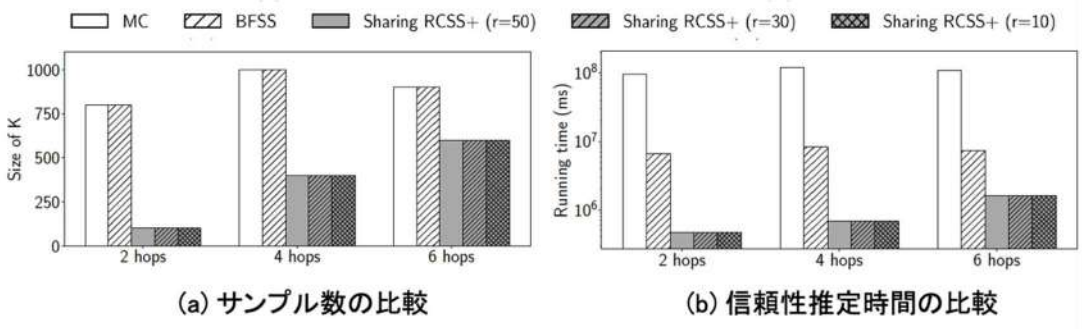

図 1. Sharing RCSS+の性能比較

- 属性付きグラフに対するコミュニティ検索の高速化

ノードに属性が付与された属性付きグラフに対するコミュニティ検索の高速化手法を開発した．コミュニティ検索は属性付きグラフおいて，ユーザが与えたクエリに対して最も適合性の

高いコミュニティを見つけ出す処理である．膨大なノードの組合せの中からコミュニティの頑健性とクエリに対する属性の類似度の最も組合せを探索する必要があり，大きな処理時間を必要とする．本研究では Data Skewness を利用した計算不要ノードの探索枝刈りを導入することで，計算時間の大幅な削減に成功した．

図 2 に提案手法（Fast enumeration）と最先端手法 LocATC [Huang and Lakshmanan, 2017]との実行時間の比較を示す．図 2 では Cornell，Texas という 2 種類の実データセットを対象にランダムに選択した 100 個のクエリを処理した際の平均実行時間を比較している．この図からもわかるように提案手法は LocATC と比較して数十倍程度高速にコミュニティを検索できていることが確認できる．

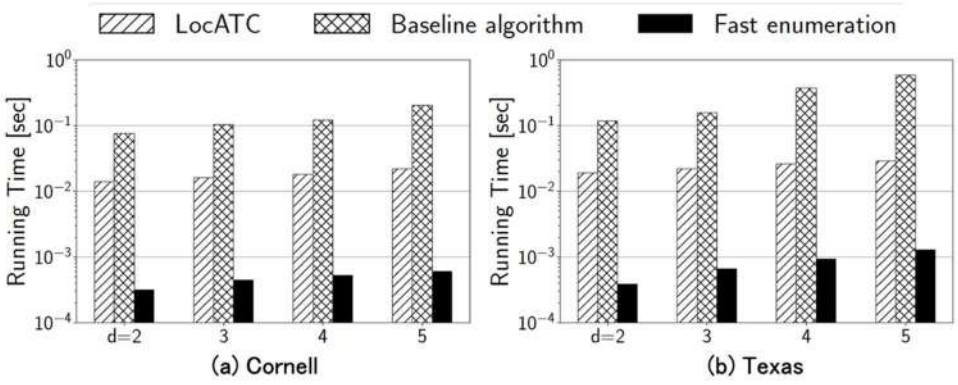

図 2. Fast enumeration 法の実行時間比較

- 分散グラフクラスタリングのスケーラビリティ向上

本研究では Data Skewness を用いた効率的な分散グラフクラスタリング DSCAN を開発した．一般的に分散グラフ処理では計算機間で同期処理を必要とする．この同期処理は一般的に大きな処理時間を要することから，同期処理を可能な限り削減することが重要な課題となる．DSCAN では Data Skewness を捉えることで同期不要なエッジを特定する．同期不要な処理を削減することで効率的な分散グラフクラスタリングを実現する．

図 3 に DSCAN の実行時間の比較を示す．代表的な既存手法よりも数十倍から数千倍程度高速な処理を実現している．また，既存の手法で処理できなかった clueweb データセットも数十秒程度で処理可能となった．また図 4 にスケーラビリティを比較した結果を示す．図 4 では各データセットに対して処理に使用した計算機台数を増加させた際の性能向上率を示しており，提案法は既存法よりも良いスケーラビリティを示していることが確認できる．

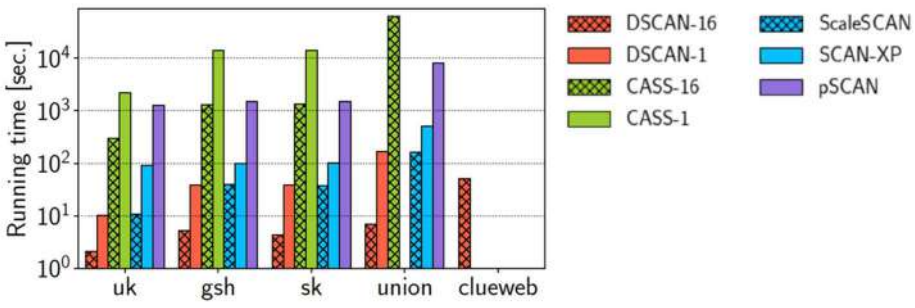

図 3 実行時間の比較

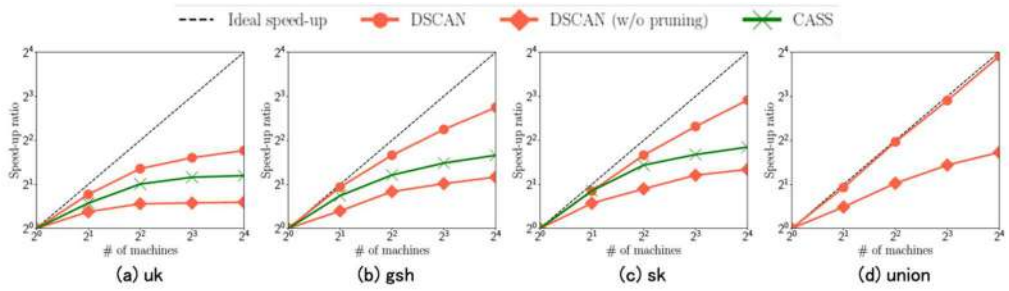

図 4. スケーラビリティの比較

【研究課題 (2)】 多次元データに対する分析アルゴリズムの高速化

本研究の目的は多次元データ処理において Data Skewness に着眼した高速化手法の有効性を検証することである. 本研究期間では全点対計算を必要とする Affinity Propagation とグラフデータベースにおける問合せ処理を題材として, 本研究のアプローチの有効性を検証した. これらの研究においても上述した研究成果と同様に, 数百万から数億件程度の実データセットにおいて, 提案法が数十倍～数百倍程度高速に計算可能であることを確認した.

- Affinity Propagation 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. 最適解に対して非常に性能の良い近似解を求めることができる性質が知られており, 幅広い領域で利用されている手法である. しかしながら, クラスタの検出には全点対計算を必要とするため膨大な計算時間を必要とする. 本研究では Data Skewness 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設計した. 本研究の提案手法 ScaleAP は数百万件程度のデータセットに対して, 従来技術よりも精度を劣化させず 100 倍以上高速な処理を可能とした.

図 5 は提案手法と代表的な最先端手法との実行時間ならびに精度の比較結果を示す. 精度の比較では, それぞれのアルゴリズムが出力した結果が Affinity Propagation とどれだけ一致しているかについて F 値を用いて比較している. この結果からもわかるように提案手法 ScaleAP は既存手法よりも 100 倍程度高速であるが, Affinity Propagation と同じクラスタリング結果を出力可能である. 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. これらの結果からも本研究が主眼をおいている Data Skewness を捉えた高速化アプローチは全点対計算を必要とするデータ処理にも応用できる可能性が示唆されたと考える.

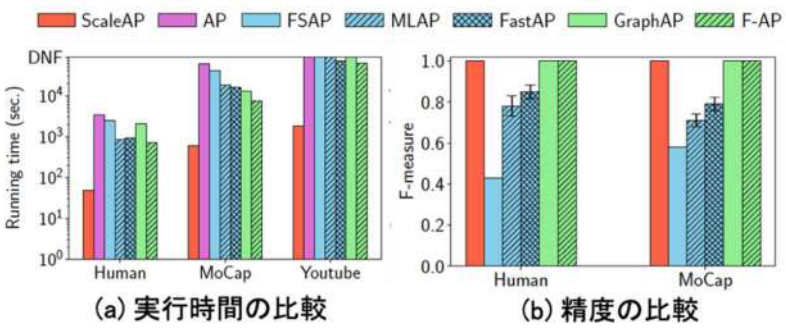

図 5. ScaleAP の性能比較

3. 今後の展開

本研究の狙いは，提案技術の開発により，誰しもが手持ちの計算環境でビッグデータ処理を出来るようにすることである．本研究は，ACT-I 期間や加速フェーズ期間を通じて，グラフデータ処理や多次元データ処理に対する Data Skewness を捉えた新しい高速化・省メモリ化技術の有効性を実験・理論の両面から確認してきた．この成果は大規模なデータを処理する際に必ずしも膨大な計算資源を必要としないことを示唆している結果であると考えている．本研究では引き続き，開発した技術のライブラリ化等を通じて，産業や社会に還元できるよう進めていく予定である．

昨今様々な IoT デバイスが普及にともなって，我々が利用可能な計算環境や計算資源の多様化が加速している．このような環境において IoT アプリケーションの高機能化を実現する上では，本研究で開発したアルゴリズムの様な高いはフォーマンスを発揮することが出来るアルゴリズムが必要不可欠になるのではないかと考えている．このような背景を受けて，本研究では IoT データなどのリアルタイムかつ計算資源の乏しい状況下におけるデータ処理の高性能化を次の研究の展開として見据えており，JST さきがけ「IoT が拓く未来」の支援のもと，IoT データ処理の高性能化に向けたプロジェクトを開始した．本プロジェクトではこれまで ACT-I を通じて開拓してきた Data Skewness を捉えた高性能化を軸として，よりリアルタイムかつマルチモーダルなデータ処理の高性能化に取り組んでいく計画である．

また，これまで ACT-I で取り組んできた研究成果の一部からも示唆されるように，Data Skewness を捉えたアプローチの有効性はアルゴリズムの高速化に限定されず，幅広い領域に展開可能なアイデアであると考えている．例えば，Data Skewness に基づきデータレイアウトを調整することで，並列計算時のメモリ(NW)バンド幅使用量を削減できることがこれまでの研究からわかってきている．これは HPC 等における HW 最適化技術にも応用可能な発見である．他にも，【研究課題(1)】で取り組んだ分散グラフクラスタリングのスケーラビリティ向上では計算機間の通信コスト削減にも本アプローチが有効であることが示唆された．これらの成果を踏まえて，今後の研究の方向性として，ACT-I を通じて開拓した技術を HW 最適化や効率的な超並列計算等のより新しい研究課題への展開していくことを見据えている．

4. 自己評価

- 研究目的の達成状況

  本研究期間を通じて，単純なグラフデータのみならず複雑な構造をもつ多様なアルゴリズムに対する Data Skewness を捉えたアプローチの構築とその有効性の検証を理論・実験の両側面行った．これは当初計画していた研究目的を十二分に達成しており，幅広いアルゴリズムに対する提案アプローチの有効性を示唆するものであったと考える．

- 研究の進め方（研究実施体制及び研究費執行状況）

  研究の実施体制及び研究費の執行状況は，新型コロナウィルス感染症の流行による海外との往来を伴う連携や出張の停止を除き，概ね計画通り進捗した．

- 研究成果の科学技術及び学術・産業・社会・文化への波及効果

  本研究期間を通じて多くのアルゴリズムに対して提案アプローチの有効性を示すことができた点は学術的に成果のあった点であると考える．特に加速フェーズ期間中はいくつかの難関国際会議にも論文が採択されており，技術的新規性や有用性，論文のイン

パクトについては高い評価を得ていると考えている．また，本研究に興味を持った海外の研究者らから問い合わせが多くあり，国際的な研究連携も進みつつある状況にある．したがって，今後着実に技術開発や研究連携を続けていくことで，学術面への波及効果は次第に大きくなると期待できる．また，主要な提案技術は論文発表と同時にソフトウェア化を行うことで産業面での波及効果を見込んでいる．

- 研究課題の独創性・挑戦性

  本研究で提案する Data Skewness を捉えたアプローチ，とりわけ Data Skewness Caching は我々が知る限り例を見ない手法である．また，上述の通り，ACT-I 期間や加速フェーズ期間を通じて複数の難関会議で発表を行い，技術的な新規性や有効性について非常に高い評価を得ており，国際的な共同研究も進みつつある．以上のことから，本研究課題の独創性・挑戦性は高いと自己評価している．

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. Hiroaki Shiokawa. Scalable Affinity Propagation for Massive Datasets. In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), February 2021 (in press)
2. Hiroaki Shiokawa, Tomokatsu Takahashi. DSCAN: Distributed Structural Graph Clustering for Billion-edge Graphs. In Proceedings of the 31st International Conference on Database and Expert Systems Applications (DEXA 2020), pp.38-54, September 2020.
3. Shohei Matsugu, Hiroaki Shiokawa, Hiroyuki Kitagawa. Fast and Accurate Community Search Algorithm for Attributed Graphs. In Proceedings of the 31st International Conference on Database and Expert Systems Applications (DEXA2020), pp.233-249, September 2020
4. Junya Yanagisawa, Hiroaki Shiokawa. Fast One-to-Many Reliability Estimation for Uncertain Graphs. In Proceedings of the 31st International Conference on Database and Expert Systems Applications (DEXA2020), pp.106-121, September 2020
5. Hiroaki Shiokawa, Toshiyuki Amagasa, Hiroyuki Kitagawa. Scaling Fine-grained Modularity Clustering for Massive Graphs. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI2019), pp.4597-4604, August 2019

（2）特許出願

研究期間累積件数：0件

（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）

- 主要な学会発表

1. Hiroaki Shiokawa, “Scalable Affinity Propagation for Massive Datasets,” the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), February 2021.
2. Hiroaki Shiokawa, Tomokatsu Takahashi, “DSCAN: Distributed Structural Graph Clustering for Billion-edge Graphs,” The 31st International Conference on Database and Expert Systems Applications (DEXA 2020), September 2020.
3. Junya Yanagisawa, Hiroaki Shiokawa, “Fast One-to-Many Reliability Estimation for

Uncertain Graphs," The 31st International Conference on Database and Expert Systems Applications (DEXA2020), September 2020.

4. Hiroaki Shiokawa, Toshiyuki Amagasa, Hiroyuki Kitagawa, "Scaling Fine-grained Modularity Clustering for Massive Graphs," The 28th International Joint Conference on Artificial Intelligence (IJCAI2019), August 2019
5. 塩川 浩昭, 天笠 俊之, 北川 博之, "基調構造を利用したグラフクラスタリングの高速化," 第 12 回データ工学と情報マネジメントに関するフォーラム (DEIM2020), March 2020.

- 受賞

1. 塩川 浩昭, 日本データベース学会 第 16 回 上林奨励賞, 2020 年 6 月 26 日
2. 塩川 浩昭, 天笠 俊之, 北川 博之, 第 12 回データ工学と情報マネジメントに関するフォーラム(DEIM2020) 優秀論文賞 "基調構造を利用したグラフクラスタリングの高速化"