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：平成２９年１０月～平成３１年３月
研究者番号：５０１４２
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グラフ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．特に 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する．決定的なアルゴリズムは入力に対して処理結果が一意になる．本研究ではこの性質を前提に，上述した Data Skewness に対する計算結果をキャッシングし，Data Skewness との同型性判定を行うことで，大規模データ処理アルゴリズムの大幅な高速化とメモリ使用量の削減を狙う．

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，本研究期間ではグラフデータに対する決定的アルゴリズムを対象とした研究開発を実施した．この研究を通じ，比較的理想的なグラフデータ並びにアルゴリズムにおいて，Data Skewness Caching が処理の高性能化に極めて有効に働くことを確認することができ

た.

具体的には，相関に基づく Modularity クラスタリングを題材として，Data Skewness Caching を導入した高速化･省メモリ化に取り組み，本研究の提案アプローチの有効性を 30 億エッジ規模の実データならびに人工データに対して評価を行った．また，ここで開発した手法の特性や性質の議論と効率化を行うために，本手法に対する理論的な解析についても合わせて実施した．Data Skewness Caching の組み込みに際して，どのような構造特徴に着目するかが性能を左右するパラメータとして存在するが，本期間では少数のノードから構成されるシンプルな構造特徴を検証の対象とした．評価結果の詳細は次節にて述べるが，大規模なデータに対して大幅な高速化性能ならびに省メモリ化性能を示すことを実験的に確認した．また，この結果を受けて，比較的シンプルなグラフモデルを前提とした Data Skewness Caching の理論的な性能限界に関する検証も実施し，提案アプローチである Data Skewness Caching は実装の効率化により高速･省メモリなデータ処理を実現できる可能性を示した．

また，本研究では上記の通り開発した提案アプローチ Data Skewness Caching を様々なグラフデータ処理に対して拡張･応用を行った．次節では，本研究で実施した拡張のひとつとして密度ベースグラフクラスタリングの高速化手法について述べる．この手法では Data Skewness Caching に加え，数億エッジ規模のグラフに出現する次数比分布に着眼した動的なノード枝刈り手法を導入した．これにより，従来のアルゴリズムと比較して大幅な高速化に成功した．

（2）詳細

- **Modularity クラスタリングの高速化･省メモリ化手法の開発**

局所相関に基づく Modularity クラスタリングの超高速･省メモリ化手法 gScarf を開発した．gScarf は Data Skewness Caching という新たなデータ処理法を提案している．このデータ処理法は，解の決定性が保証されるシンプルな部分グラフ構造の計算結果を動的にメモ化することで，計算に必要な時間･空間コストの削減を図る．

提案手法 gScarf の実行速度を最先端の手法である CorMod [Duan et al., KDD’15]と比較を行い，約 300 万エッジ (YT) から約 30 億エッジ (TW) 規模のデータに対して，最大で 1,000 倍以上の高速化性能を確認した．YT データにおける各手法の計算回数のヒストグラムを図 2 に示す．図 2 の各ピクセルはグラフの各エッジを示し，エッジの計算回数が多いほど対応するピクセルが黄色にハイライトされている．図 2 からわかるように，gScarf は CorMod と比較して計算回数を 1%未満にまで削減することに成功している．

また，提案手法 gScarf は実データに対してメモリ使用量を従来の約 30%程度まで抑制するとともに，その精度は CorMod と同程度になることを実験的に確認している．さらに本研究では，Data Skewness Caching の実行速度･メモリ消費量に関する有効性を理論的に解析し，理想的なグラフに対して時間･空間計算量をグラフサイズの数千分の一まで抑えられることを示した．

- **密度ベースグラフクラスタリングの高速化手法の開発**

密度ベースクラスタリング SCAN の超高速化手法 ScaleSCAN [1]を開発した．我々は極め

て大規模なグラフのみにおいて隣接ノード間の次数比が大きく偏るという性質を新たに発見した．この発見に基づき，ScaleSCAN は上述した Data Skewness Caching に加えて，次数比に応じた動的な計算ノード枝刈り手法を導入している．

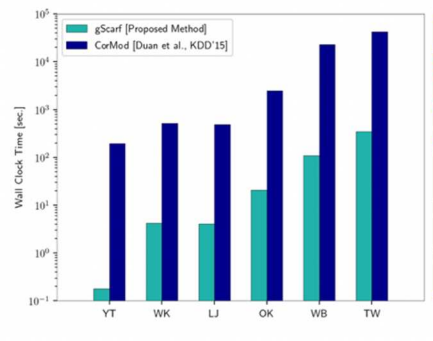

図 1. 実行時間の比較

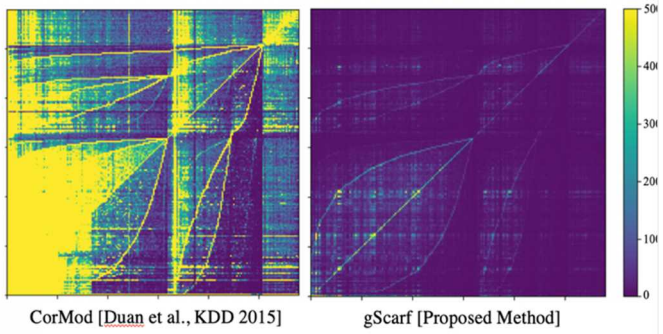

図 2. 計算回数のヒストグラム (TW データ)

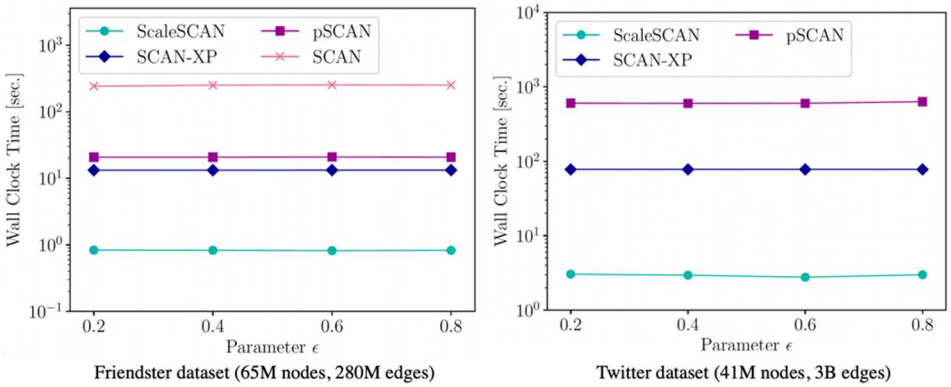

図 3. 実行時間の比較

提案手法 ScaleSCAN の実行速度を，最先端の手法 SCAN-XP，pSCAN ならびにベースライン手法 SCAN と比較を行い，500 倍から 1,000 倍程度の高速化性能を確認した (図 3)．特に，Twitter dataset では，SCAN は 24 時間以内に処理が終了しなかったのに対し，提案手法 ScaleSCAN は約 6.4 秒で厳密解を計算可能であることを確認した．

3. 今後の展開

本研究提案の狙いは，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 期間においてグラフデータに対する 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．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．

更に 2 年間の加速フェーズでは，この技術･発見をグラフデータに限定されず，多様なデータとアルゴリズムに応用することに挑戦する．本研究を通じた将来の見通しとして，Data Skewness という独自の着眼点を基にした様々なデータ処理技術の再設計を行い，高性能な解法群の構築と

体系化を目指したいと考えている.

4. 自己評価

- 研究目的の達成状況

ACT-I 期間の研究を通じて, 単純な (重み無し・無向) グラフ構造を対象とした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, 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

ACT-I 期間中は難関会議に投稿した論文の多くが不再録となり波及効果は限定的であったが, いずれの査読結果でも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. したがって, 今後着実に論文出版のプロセス (特に難関会議への挑戦) 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 期間を通じて複数の難関会議に投稿を行い, 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いる. 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, Tomokatsu Takahashi, Hiroyuki Kitagawa. ScaleSCAN: Scalable Density-based Graph Clustering. The 29th International Conference on Database and Expert Systems Applications. 2018, pp. 18-34
2. Tomoki Sato, Hiroaki Shiokawa, Yuto Yamaguchi, Hiroyuki Kitagawa. FORank: Fast ObjectRank for Large Heterogeneous Graphs. The Web Conference 2018. 2018, pp.103-104 (poster)
3. Hiroaki Shiokawa, Yasunori Futamura. Graph Clustering via Cohesiveness-aware Vector Partitioning. The 20th International Conference on Information Integration and Web-based Applications and Services. 2018, pp. 33-40.

(2) 特許出願

研究期間累積件数: 0件

(2) 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. 佐藤 朋紀, 塩川 浩昭, 北川 博之,"グラフの構造情報を用いた ObjectRank の高速化," 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM2019), March 2019.
2. 松下 朋弘, 塩川 浩昭, 北川 博之,"メッセージ集約に基づく Affinity Propagation の高速化," 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM2019), March 2019.
3. 山崎 耕太郎, 塩川 浩昭, 北川 博之,"クラスタの収束性を用いた逐次的枝刈りによる RankClus の高速化," 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM2019), March 2019.
4. 真次 彰平, 塩川 浩昭, 北川 博之,"属性付きグラフに対するビームサーチを用いたコミュニティ検索," 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM2019), March 2019.
5. 佐藤 朋紀, 塩川 浩昭, 北川 博之,"選択的重要度先読みを用いた ObjectRank の高速化," 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), March 2018.

- 受賞

1. 塩川 浩昭, 情報処理学会 2018 年度山下記念研究賞, 2019 年 3 月 15 日
2. 佐藤 朋紀, 塩川 浩昭, 北川 博之, 第 10 回データ工学と情報マネジメントに関するフォーラム, 優秀論文賞, 2018 年 6 月 22 日