ビッグデータのための並列アルゴリズムの開発

Parallel Algorithms for BigData Processing

塩川 浩昭
筑波大学 計算科学研究センター

１．研究目的

ビッグデータの高速分析は近年高い注目を集めている技術領域である．我々のプロジェクトでは実世界の極めて大規模なビッグデータを対象に，並列計算環境を利用した高速なアルゴリズムの開発している．我々は過去 2 年間の筑波大学学際共同利用にて，数億ノード規模のグラフに対する高速分析アルゴリズム *SCAN-XP* ならびに *DSCAN* と数億レコードのテキストを対象とした集合間類似結合手法開発した．本プロジェクトの期間では，これらの手法を用いた実データに対する大規模な性能評価を目的とする．具体的には，昨年度開発した DSCAN および集合間類似結合手法を対象として，複数の Intel Xeon Phi を活用して数 TB 規模のデータを対象とした性能評価を行うとともに，各アルゴリズムのパフォーマンスチューニングを行う．

２．研究成果の内容

本年度の主要な研究成果は，(1)大規模なグラフデータを対象とした分散並列構造的クラスタリング DSCAN の開発･性能検証ならびに (2)大規模なデータベースを対象とした集合間類似結合処理の並列化に関する性能評価の 2 点である．以下のそれぞれの概要を述べる．

(1) 大規模なグラフデータに対する分散並列構造的クラスタリング DSCAN

データの関連性を表現するグラフデータに対するクラスタ分析技術技術の高速化に向けて，本研究期間では，分散並列構造的クラスタリング DSCAN の開発・性能評価ならびにチューニングに取り組んだ．DSCAN の高速化では，複数の Intel Xeon Phi 間における通信コストを削減する枝刈りにより，従来技術と比較して大幅な計算時間削減を達成した．

詳細な従来技術との実行速度の比較を図 1 に示す．図 1 に示した DSCAN は本研究の提案手法であり，ScaleSCAN ならびに pSCAN-XP は既存の並列処理手法，pSCAN は並列化を用いない高速化手法である．図の x 軸は評価に用いたデータセットを表しており，uk, gsh, sk, union はそれぞれ 9.3 億，18 億，19 億，55 億エッジ規模のグラフデータである．図 1 からも分かるように本研究で提案した DSCAN は既存手法の中で最も高速であり，概して 5 倍～250 倍程度の高速化に成功している．特に，最も巨大な 55 億エッジ規模のグラフデータ union では提案手法 DSCAN の

みが実行結果を返すことが可能であり，他の手法では 24 時間以内に解を得ることが出来なかった．具体的には DSCAN は union データセットを約 9.21 秒で処理可能であることを示した．以降の評価より提案手法の有効性を本研究期間では示すことが出来た．

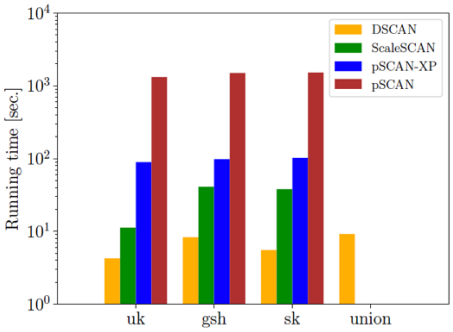

図 1. DSCAN の実行時間比較

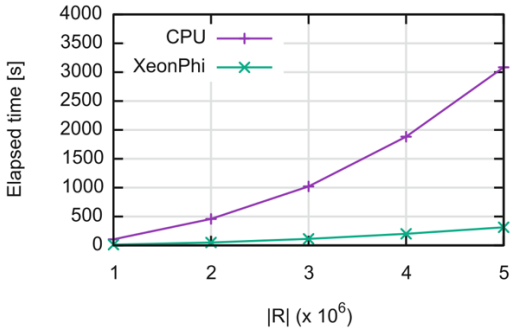

図 2. 集合間類似結合の実行時間比較

(2) 大規模なデータベースを対象とした集合間類似結合処理

集合間類似結合とは，集合をレコードとする二つのレコード集合から閾値以上の類似度を示すレコードのペアを抽出・列挙する基本的なデータ処理である．本研究期間では，MinHash 法による Locality Sensitive Hashing を用いた類似結合処理を Intel Xeon Phi に最適化することで高速化を図った．

実際のテキストデータベースを用いた性能評価実験結果を図 2 に示す．図 2 ではデータセットのサイズを変化させた際の CPU を用いた集合間類似結合計算と，本研究で提案する Intel Xeon Phi を用いた計算方法との実行時間の比較を行っている．図からもわかるように，CPU を用いた場合，データセットサイズの増加に対して指数関数的に計算時間が増加するのに対し，提案手法（XeonPhi）は一貫して高速であり，ほぼ線形のスケーラビリティを示している．以降の評価より提案手法の有効性を本研究期間では示すことが出来た．

３．学際共同利用として実施した意義

近年，計算情報学の分野では利用可能なデータの規模が増加の一途をたどっており，学際共同利用を通じて提供される高性能な計算環境無くしては処理できない状況となっている．したがって，学際共同利用として実施した意義が大きい．

４．今後の展望

今後はまず，より大規模かつ多様な種類のデータに対して，本年度開発した成果の性能検証を深める．また，本研究を通じて獲得した並列化や高速化の手法を他のアルゴリズムへと適用し，幅広い分析の高速化に発展させる予定である．

５．成果発表

(1) 学術論文

- Hiroaki Shiokawa, Yasunori Futamura, “Graph Clustering via Cohesiveness-aware Vector Partitioning,” In Proc. iiWAS 2018, pp.33-40, 2018（査読あり）
- Tomohiro Matsushita, Hiroaki Shiokawa, Hiroyuki Kitagawa, “C-AP: Cell-based Algorithm for Efficient Affinity Propagation,” In Proc. iiWAS 2018, pp. 156-163, 2018（査読あり）
- Kotaro Yamazaki, Tomoki Sato, Hiroaki Shiokawa, Hiroyuki Kitagawa, “Fast Algorithm for Integrating Clustering with Ranking on Heterogeneous Graphs,” In Proc. iiWAS 2018, pp. 24-32, 2018（査読あり）
- Hiroaki Shiokawa, Tomokatsu Takahashi, Hiroyuki Kitagawa, “ScaleSCAN: Scalable Density-based Graph Clustering,” In Proc. DEXA 2018, pp. 18-34, 2018（査読あり）
- Tomoki Sato, Hiroaki Shiokawa, Yuto Yamaguchi, Hiroyuki Kitagawa, “FORank: Fast ObjectRank for Large Heterogeneous Graphs,” In Proc. WWW, pp. 103-104, 2018（査読あり）

(2) 学会発表

- 真次 彰平, 塩川 浩昭, 北川 博之, “属性付きグラフに対する効率的なコミュニティ問合せ処理,” 情報処理学会 第 81 回全国大会, 2019 年 3 月 14 日~3 月 16 日，福岡大学
- 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, 2019 年 3 月 4 日～3 月 6 日，ホテルオークラ JR ハウステンボス
- 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, 2019 年 3 月 4 日～3 月 6 日，ホテルオークラ JR ハウステンボス
- 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, 2019 年 3 月 4 日～3 月 6 日，ホテルオークラ JR ハウステンボス
- 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, 2019 年 3 月 4 日～3 月 6 日，ホテルオークラ JR ハウステンボス

(3) その他

該当なし

使用計算機	使用計算機に○	配分リソース※	
		当初配分	追加配分
COMA			
Oakforest-PACS	○	4,800	

※配分リソースについてはノード時間積をご記入ください。