受付 ID	17a2
分野	計算情報学

大規模データ分析手法の高速化に関する研究

Fast Algorithms for Big Data Analysis

塩川 浩昭
筑波大学 計算科学研究センター

１．研究目的

本研究期間では数億～数十億件のレコードから構成される実世界の極めて大規模なグラフやテキストを含むデータを対象に，並列計算環境を利用した高速な分析アルゴリズムの開発に取り組む．我々は平成 28 年度の筑波大学学際共同利用プロジェクトにて，数億ノード規模のグラフに対する高速な分析アルゴリズム *SCAN-XP* を開発した．本研究期間では，昨年度の研究成果を基により大規模なデータを対象とした高速なデータ分析手法を開発する．

２．研究成果の内容

本年度は，(1)大規模なグラフデータを対象とした構造的クラスタリングならび ObjectRank 解析，および(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して研究成果を得た．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構造的クラスタリングならび ObjectRank 解析

データの関連性を分析するグラフデータ解析技術の高速化に向けて，本研究期間では，構造的クラスタリング SCAN ならびに ObjectRank の高速化に取り組んだ．SCAN 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，従来処理できなかった 60 億ノード規模のクラスタリングを世界で初めて実現した．また，ObjectRank の高速化では，疎行列ベクトル積において早期に値の収束が見込めるノードの推定手法を理論的に導出し，これにより従来技術の 10 倍以上の高速化が見込めることを実験的に確認した．

(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．テキストデータを用いた実験により，提案手法は既存手法より約 40 倍程度高速であることを確認した．

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

- Tomoki Sato, Hiroaki Shiokawa, Yuto Yamaguchi, Hiroyuki Kitagawa, “FORank: Fast ObjectRank for Large Heterogeneous Graphs,” In Proc. WWW2018, 2018. （査読あり，印刷中）
- 高橋 知克, 塩川 浩昭, 北川 博之,"メニーコアプロセッサを用いた構造的類似度に基づくグラフクラスタリングの高速化," 情報処理学会論文誌：データベース (TOD76), Vol.10, No.4, pp.1-5, December 2017.（査読あり）
- 佐藤 朋紀, 塩川 浩昭, 山口 祐人, 北川博之,"大規模グラフに対する ObjectRank の高速な近似 Top-k 検索," 情報処理学会論文誌：データベース (TOD76), Vol.10, No.4, pp.11-15, December 2017.（査読あり）

(2) 学会発表

- 菅野 健太, 天笠 俊之, 北川 博之, “メニーコアプロセッサを用いた大規模な集合間類似結合の高速化,” 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 (DEIM 2018) , H6-2, 2018 年 3 月 4 日～3 月 6 日.
- 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 (DEIM 2018) , I7-5, 2018 年 3 月 4 日～3 月 6 日.
- 高橋 知克, 塩川 浩昭, 北川 博之, “Intel Xeon Phi による SCAN クラスタリングの分散並列化,” 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 (DEIM 2018) , H6-1, 2018 年 3 月 4 日～3 月 6 日.

(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
HA-PACS/TCA			
COMA	○	3200	
Oakforest-PACS	○	4800	
※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。			