受付 ID	16a7
分野	計算情報学

大規模グラフ分析アルゴリズムの高速化に関する研究

Fast Algorithm for Large-scale Graph Analysis

塩川 浩昭
筑波大学 計算科学研究センター

１．研究目的

本課題の目的は，数億ノードから構成される実世界の極めて大規模なグラフを対象に，超並列計算環境を利用した高速な分析アルゴリズムを開発することである．特に今年度は，本目的の達成に向けて，これまで我々が計算情報学の視点から開発を行ってきたグラフクラスタリングソフトウェアを題材に，Intel Xeon Phi を活用したよりスケーラビリティの高いアルゴリズムの実現を目標とし，研究活動を遂行した．

２．研究成果の内容

本研究課題では，構造的類似度に基づくクラスタリングの超並列化手法を開発したグラフクラスタリングはグラフの中に存在するコミュニティ構造を理解する上で重要な要素技術である．その中でも構造的類似度に基づくクラスタリング手法 SCAN は高い精度でクラスタを検出することが出来ることから幅広いアプリケーションで利用されている．しかしながら，SCAN はグラフに含まれる全てのノードとエッジに対して構造的類似度計算を行う必要があり，大規模なグラフを対象とした場合に計算量が爆発するという問題がある．この問題に対して，これまでいくつかの高速化手法が提案されてきた．我々もこれまで SCAN++と呼ばれる高速化手法を提案してきたが，SCAN++を以ってしても Web やソーシャルネットワークのような数億ノードから数十億ノード規模のグラフを対象とした場合に，依然として十数時間から数週間程度の計算時間を要するという課題があった．

本研究課題では，上述した課題を解決するために，COMA が搭載している Intel Xeon Phi Co-processor を活用した SCAN の超並列化手法 SCAN-XP を開発した．SCAN-XP では，実世界に存在するグラフには次数分布の偏りや 3 部クリーク構造が頻出すると行った構造特性を持つことに着眼し，(1) 実世界のグラフの構造特性による並列化性能向上のボトルネックを解消するデータレイアウトを与えるとともに，(2)Intel Xeon Phi Co-processor の持つ多くの物理コアと 512 ビット SIMD 演算を最大限に活用するためのアルゴリズムの最適化を行った．性能評価のために大規模な実データを用いた実験を実施し，提案手法 SCAN-XP は既存手法である SCAN と比較して，分析時間が 100 倍以上高速化されていることを確認した．また，1 億ノード規模

のグラフに対するクラスタリングも，SCAN-XP を用いることで 30 秒程度の時間で実行することが可能であり，この処理性能は我々が知る限り 2017 年 4 月現在で世界最高性能である．

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，汎用の CPU を用いた分析処理の並列化のみではデータを処理しきれない現状に直面している．特に本研究課題で対象とするデータセットは 1.2TB～2.5TB と大きく，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．本研究課題を通じて，従来現実的な計算時間では処理できなかった規模のデータ分析を世界に先駆けて実現した．この成果は，計算情報学分野における研究活動の進展に対して，非常に大きな意義がある成果であると考える．

４．今後の展望

今後は本年度開発した成果を基に，より大規模かつ多様な種類のデータを分析可能なアルゴリズムの開発へと発展させる計画である．具体的には，本年度対象としたデータ規模の 10 倍～100 倍規模のデータや，これまで対象としてこなかったテキストデータなどを扱ったアルゴリズムの開発を進めている．

５．成果発表

(1) 学術論文

- Tomokatsu Takahashi, Hiroaki Shiokawa, Hiroyuki Kitagawa, "SCAN-XP: Parallel Structural Graph Clustering Algorithm on Intel Xeon Phi Coprocessors," In Proceedings of the 2nd ACM SIGMOD Workshop on Network Data Analytics (NDA 2017), Chicago IL USA, May 19th 2017, （査読有, 印刷中）

(2) 学会発表

- 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"，第 9 回データ工学と情報マネジメントに関するフォーラム(DEIM2017)，E2-2，岐阜県高山市，2017 年 3 月 6 日（口頭発表・ポスター）

(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※
HA-PACS		
HA-PACS/TCA		
COMA	○	112.5

※配分リソースについては 32node 換算時間をご記入ください。