様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １４ 日現在

機関番号：１２１０２
研究種目：若手研究
研究期間：2018～2021
課題番号：１８Ｋ１８０５７

研究課題名（和文）超並列計算環境における大規模グラフの実時間問合せ処理

研究課題名（英文）Real Time Graph Query Processing on Massively Parallel Environment

研究代表者
塩川　浩昭（Shiokawa, Hiroaki）

筑波大学・計算科学研究センター・准教授

研究者番号：９０７７５２４８

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では実世界のグラフデータが持つ性質に着眼したグラフ並列処理技術を開発し，大規模グラフに対する多様な問合せ処理問題の高速化手法を開発した．具体的には，(1)大規模重みなし無向グラフに対するコミュニティ問合せ，(2)ヘテロジニアスグラフにおける類似問合せ，(3)属性付きグラフにおける密部分グラフ問合せ，(4)複雑ネットワークに対する範囲問合せについて取り組み，分散並列環境における高速・スケーラブルなアルゴリズム群を開発した．(1)から(4)までにいずれに研究成果においても従来手法と同程度の問合せ処理品質を保証しつつ，10倍から10,000倍程度の高速化を実現した．

研究成果の学術的意義や社会的意義
本研究が対象とした大規模グラフデータはタンパク質データや化学物質データなどを扱う医療・ライフサイエンス分野，ソーシャルネットワークやWebデータなどを扱うSNSアプリケーションなど，我々の日常生活に密接に関わるデータである．本研究が開発した高速・高精度・スケーラブルな問合せ処理アルゴリズムはこれらの利用場面において，研究や開発サイクルの加速やアプリケーションの利便性向上に大きく寄与する技術である．

研究成果の概要（英文）：In this research project, we have developed graph parallel processing techniques that focus on the properties of real-world graph data, and have developed methods for accelerating various query processing problems for large-scale graphs. Specifically, we addressed (1) community search queries for large unweighted undirected graphs, (2) similarity search queries for heterogeneous graphs, (3) dense subgraph search queries for attributed graphs, and (4) range queries for complex networks, and we have developed a set of fast and scalable algorithms for distributed parallel environments. All of the results from (1) to (4) are 10 to 10,000 times faster than conventional methods while guaranteeing the same level of query processing quality.

研究分野：データベース

キーワード：データベース　グラフ　問合せ処理　ビッグデータ

１．研究開始当初の背景

時々刻々と生成される大規模グラフ構造を高速・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．大規模データに対して，高速かつ高精度に処理結果を獲得するためには超並列計算環境を用いることが一般的である．しかし，グラフ分析処理は一般的にノード間の接続傾向を分析する必要があることから，大きな同期処理のオーバヘッド等が生じてスケーラビリティが低下する．この性質は様々なグラフ分析処理の中でも特にグラフに対する問合せ処理に対して顕著に現れ，現時点ではグラフの問合せ処理に対する効率的な並列処理手法が確立されていない．

２．研究の目的

本研究の目的は超並列処理環境における大規模グラフ問合せ処理の高速化である．大規模なグラフに対する既存の並列化手法では，並列度に対してスケーラビリティが低下する場合が多い．ゆえに，超並列計算環境が持つ潜在的な演算性能と現状の大規模グラフにおける問合せ処理性能のギャップを埋める技術の開発が重要となる．本研究ではグラフ問合せ処理の並列化オーバヘッドの抑制を図ることで，超並列処理環境におけるグラフ問合せ処理の高速化を狙う．

３．研究の方法

本研究期間では大別して【研究項目 1】グラフ並列処理におけるプリミティブなデータ操作の最適化，【研究項目 2】プリミティブなデータ操作に基づくグラフ問合せ処理アルゴリズムの開発に取組む．【研究項目 1】では本研究の基盤となるグラフ並列処理プリミティブの最適化を行う．具体的には，グラフ分割に伴う同期処理オーバヘッドの抑制手法，SIMD 命令を用いた効率的なデータ探索・データ処理手法等の開発を行う．【研究項目 2】では研究項目 1 で開発したグラフ問合せ処理プリミティブに基づいて大規模グラフにおける具体的な問合せ処理手法の高速化アルゴリズムを開発する．

４．研究成果

(1) 分散並列環境におけるグラフコミュニティ問合せの高速化

本研究では次数比の偏りを用いた効率的な分散コミュニティ問合せ法 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．提案手法 DSCAN では次数比の偏りを捉えることで同期不要なエッジを特定することで不要な同期処理を削減し，効率的な分散コミュニティ問合せを実現する．図 1 に DSCAN の実行時間の比較を示す．図 1 の DSCAN-16 ならびに CASS-16 は 16 台の計算機を用いて分散並列処理した際の実行時間，DSCAN-1 ならびに CASS-1 は 1 台の計算機を用いて処理した際の実行時間を表している．この結果からも，代表的な既存手法よりも数十倍から数百倍程度高速な処理を実現している．また，既存の分散処理手法（CASS-1，CASS-16）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ている．また図 2 にスケーラビリティを比較した結果を示す．図 2 では 1 台の計算機で実行した際の計算速度に対する性能向上率を示している．DSCAN (w/o pruning) は通信コスト削減を行わない場合の DSCAN である．この結果からも，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．

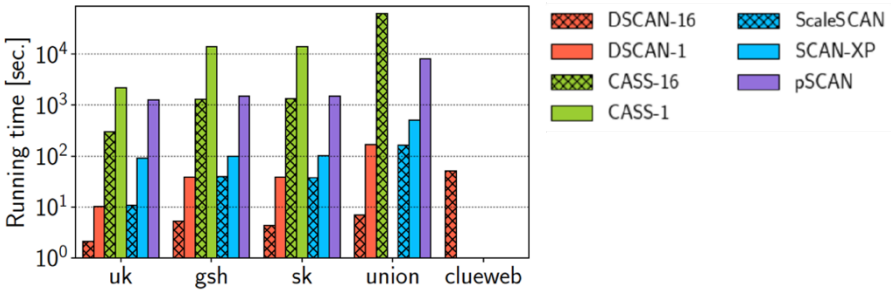

図 1. 実行時間の比較

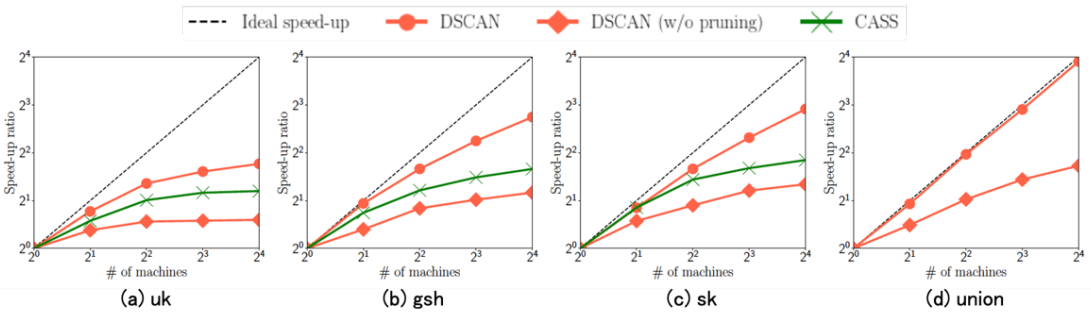

図 2. スケーラビリティの比較

(2) ヘテロジニアスグラフにおける類似コミュニティ問合せの高速化

本研究ではヘテロジニアスグラフを対象とした類似コミュニティ問合せ処理の高速化手法を提案した．複数種類のノードから構成されるヘテロジニアスグラフに対して，ユーザがクエリノードを指定すると，提案手法はそのクエリに対して類似度の高いコミュニティ上位 k 件をグラフ中から高速に検出することができる．また，提案手法はヘテロジニアスグラフのノードが追加・削除されるような場合においても誤差逆伝播に基づく検索結果の差分更新計算を並列に実行することで，グラフの動的な変化に対しても高速に問合せ応答することができる．

図 3 において提案手法（Proposed method）と従来手法（RankClus）の実行時間の比較結果を示す．図 3（左図）では両手法を 1 スレッドで実行した際の実行時間を比較している．この図では検出するコミュニティ数を K とし，K の値を変化させたときの実行時間の推移も比較している．この結果からも明らかな通り，提案手法は 1 スレッドにおいても従来手法の約 2 倍程度の高速化率を達成していることがわかる．また，図 3（右図）では提案手法においてスレッド数 T を大きくした際の高速化率を示しており，スレッド数を十分に大きく設定した場合，従来手法に対して最大で 4 倍程度の高速化が実現できていることがわかる．図 4 ではグラフにノードの追加・削除が発生した際の提案手法の問合せ処理近似性能を示している．図 4 は提案手法が用いる誤差逆伝播法の停止条件として指定するパラメータϵの設定値を横軸とし，その設定値に対する NMI の値を描画している．この結果からも分かる通り，提案手法はいかなる設定においても NMI の値が 0.9 以上となっており，高い近似性能を持っていることが示唆される．

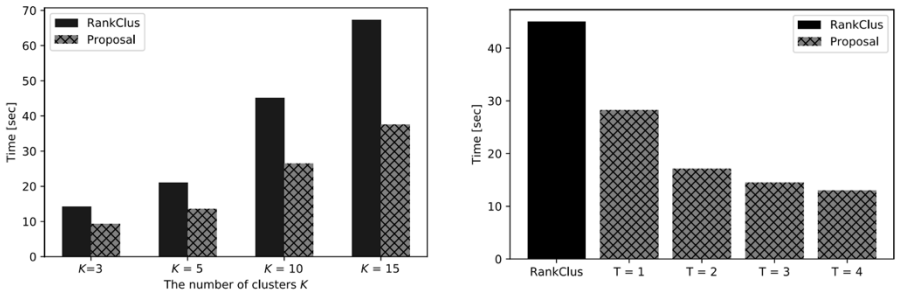

図 3. 1 スレッドでの実行時間の比較（左図）と並列度に対する実行時間の比較（右図）

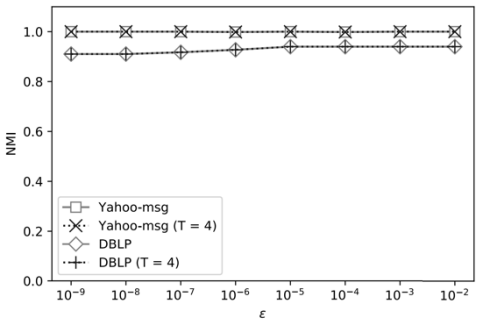

図 4. 提案手法における NMI の比較

(3) 属性付きグラフにおける密部分グラフ問合せの高速化

本研究ではグラフ中の各ノードにノードを特徴づける属性が複数与えられた属性付きグラフを対象とした問合せ処理について考える．ユーザがこのグラフの中からクエリとなるノードと属性集合を指定したとき，クエリに対して最も属性類似度が高く，クエリから距離の近い密部分グラフを検出する問題を対象とする．この問題に対する従来研究では(k,d)-truss と呼ばれるクリークを緩和した密部分グラフモデルを想定し，ユーザが与えたパラメータ k と d に対して，グ

ラフ内に存在するすべての(k, d)-truss のうち属性類似度が最大となるものを見つける．しかし，このアプローチでは多様な構造をもつ実世界のグラフにおいて密部分グラフを検出することができない場合が多く存在する．そこで，まず本研究では(k, d)-truss の構造制約を緩和した Flexible ATC 問題を定義した．Flexible ATC 問題ではグラフ中の任意の k を持つ(k, d)-truss を対象として，属性値が最大となる密部分グラフを探索する．この探索計算は従来研究と比較しても多くの計算時間を必要とするため，本研究ではさらに高速な問合せ処理アルゴリズムを提案することで，従来研究よりも高速に高品質な密部分グラフ検出を実現した．具体的には，約 110 万ノード・300 万エッジの属性付きグラフから，従来手法よりも F1 値の高い密部分グラフを 0.1 秒程度で検出することが可能である．

図 5 において提案手法（Fast Enumeration）と先行研究（LocATC）の実行時間ならびに F1 値の比較結果を示す．左図からも明らかなように提案手法は先行研究と比較して，概ね 10 倍から 100 倍程度高速に問合せ処理を実行することが可能である．また右図が示すように，提案手法が検出する密部分グラフは高い F1 値を示すことが実験から明らかになっている．これは上述した(k, d)-truss の構造制約緩和によるものである．提案手法は任意の k の設定値に対して密部分グラフを検索することができるため，事前に k の値を設定する必要がある先行研究よりも柔軟な構造をもつ部分グラフを探索することが可能である．その結果として，先行研究よりも高い F1 値を示すことができている．

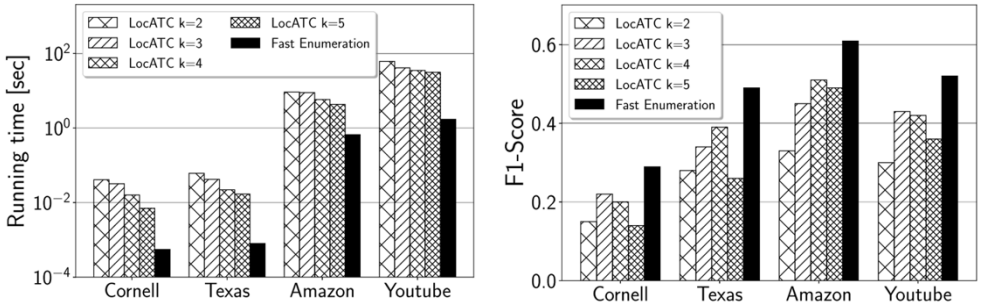

図 5. 実行時間の比較（左図）と F1 値の比較（右図）

(4) グラフ k 最近傍問合せの高速化

本研究では複雑ネットワークにおいて高速に k 最近傍検索を行うための索引構築手法を提案した．本研究では複雑ネットワークに含まれる頻出構造を事前抽出し，それらに対して個別の索引を構築することで，複雑ネットワークに対する索引構築と索引を用いた k 最近傍検索の高速化を実現する．本論文の提案手法は約 2,200 万エッジから構成される複雑ネットワークに対する索引構築を 5 秒未満，k 最近傍検索時間を 0.5 秒程度にまで高速化することに成功した．

図 6 において提案手法 (Proposed method)の索引構築時間と k 最近傍検索時間を最先端手法(G-Tree, ILBR) と比較した結果を示す．図中の CAL, NY, FLA, および TV は平面グラフであり，それ以外は複雑ネットワークである．提案手法はいずれのデータセットにおいても比較手法よりも高速な索引構築と k 最近傍検索を実現していることがわかる．また，平面グラフと複雑ネットワークにおける提案手法の高速化率を比較すると，複雑ネットワークにおいて提案手法は大幅な高速化を達成している．

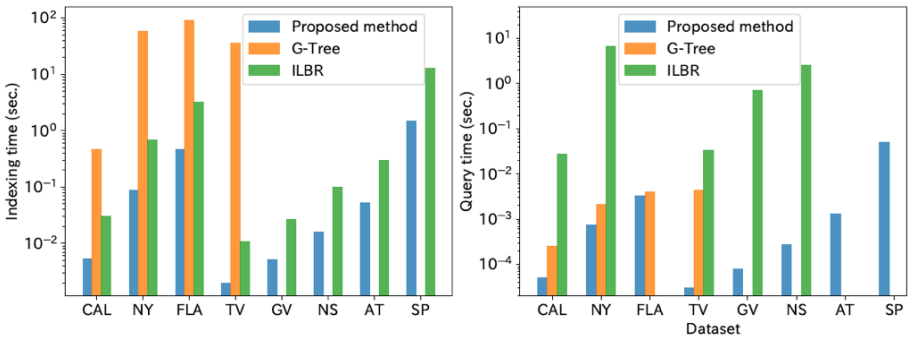

図 6. 索引構築時間の比較（左図）と k 最近傍検索時間の比較（右図）

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Matsugu Shohei、Shiokawa Hiroaki、Kitagawa Hiroyuki	４．巻 29
２．論文標題 Fast Algorithm for Attributed Community Search	５．発行年 2021年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 188～196
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.29.188	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yamazaki Kotaro、Matsugu Shohei、Shiokawa Hiroaki、Kitagawa Hiroyuki	４．巻 28
２．論文標題 Fast and Parallel RankClus Algorithm based on Dynamic Rank Score Tracking	５．発行年 2020年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 453～461
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.28.453	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計23件（うち招待講演　0件／うち国際学会　10件）

１．発表者名 Junhu Wang, Shikha Anirban, Toshiyuki Amagasa, Hiroaki Shiokawa, Zhiguo Gong, Md. Saiful Islam
２．発表標題 A Hybrid Index for Exact Shortest Distance Queries
３．学会等名 the 21st International Conference on Web Information Systems Engineering (WISE2020)（国際学会）
４．発表年 2020年

１．発表者名 Hiroaki Shiokawa, Tomokatsu Takahashi
２．発表標題 DSCAN: Distributed Structural Graph Clustering for Billion-edge Graphs
３．学会等名 the 31st International Conference on Database and Expert Systems Applications (DEXA2020)（国際学会）
４．発表年 2020年

１．発表者名
Shohei Matsugu, Hiroaki Shiokawa, Hiroyuki Kitagawa

２．発表標題
Fast and Accurate Community Search Algorithm for Attributed Graphs

３．学会等名
the 31st International Conference on Database and Expert Systems Applications (DEXA2020)（国際学会）

４．発表年
2020年

１．発表者名
Junya Yanagisawa, Hiroaki Shiokawa

２．発表標題
Fast One-to-Many Reliability Estimation for Uncertain Graphs

３．学会等名
the 31st International Conference on Database and Expert Systems Applications (DEXA2020)（国際学会）

４．発表年
2020年

１．発表者名
小林 瑞季, 真次 彰平, 塩川 浩昭

２．発表標題
グラフk近傍検索のための高速な索引構築手法の提案

３．学会等名
情報処理学会 第83回全国大会

４．発表年
2021年

１．発表者名
真次 彰平, 塩川 浩昭

２．発表標題
高速な最大k-Plex探索アルゴリズムの提案

３．学会等名
第13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 2021)

４．発表年
2021年

１．発表者名
柳澤 隼也，塩川 浩昭

２．発表標題
グラフ構造に基づく信頼性クエリの高速推定

３．学会等名
第13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 2021)

４．発表年
2021年

１．発表者名
小林 瑞季，真次 彰平，塩川 浩昭

２．発表標題
グラフに対する効率的なk最近傍検索のための索引構築手法

３．学会等名
第13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 2021)

４．発表年
2021年

１．発表者名
Kotaro Yamazaki, Shohei Matsugu, Hiroaki Shiokawa, Hiroyuki Kitagawa

２．発表標題
Fast and Parallel RankClus Algorithm based on Dynamic Rank Score Tracking

３．学会等名
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

４．発表年
2020年

１．発表者名
真次 彰平，塩川 浩昭，北川 博之

２．発表標題
属性付きコミュニティ検索におけるビームサーチの高速化

３．学会等名
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

４．発表年
2020年

１．発表者名
柳澤 隼也，塩川 浩昭

２．発表標題
1対全ノードに対するs-t信頼性の高速推定

３．学会等名
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

４．発表年
2020年

１．発表者名
柳澤 隼也，塩川 浩昭

２．発表標題
層化抽出法を用いた1対全ノードに対するs-t信頼性推定の高速化

３．学会等名
情報処理学会 第82回全国大会

４．発表年
2020年

１．発表者名
Kotaro Yamazaki, Shohei Matsugu, Hiroaki Shiokawa, Hiroyuki Kitagawa

２．発表標題
Fast RankClus Algorithm via Dynamic Rank Score Tracking on Bi-type Information Networks

３．学会等名
Proceedings of the 21st International Conference on Information Integration and Web-based Applications and Services（国際学会）

４．発表年
2019年

１．発表者名
Shohei Matsugu, Hiroaki Shiokawa, Hiroyuki Kitagawa

２．発表標題
Flexible Community Search Algorithm on Attributed Graphs

３．学会等名
Proceedings of the 21st International Conference on Information Integration and Web-based Applications and Services（国際学会）

４．発表年
2019年

１．発表者名
Hiroaki Shiokawa, Yasunori Futamura

２．発表標題
Graph Clustering via Cohesiveness-aware Vector Partitioning

３．学会等名
the 20th International Conference on Information Integration and Web-based Applications and Services（国際学会）

４．発表年
2018年

１．発表者名
Tomohiro Matsushita, Hiroaki Shiokawa, Hiroyuki Kitagawa

２．発表標題
C-AP: Cell-based Algorithm for Efficient Affinity Propagation

３．学会等名
the 20th International Conference on Information Integration and Web-based Applications and Services（国際学会）

４．発表年
2018年

１．発表者名
Kotaro Yamazaki, Tomoki Sato, Hiroaki Shiokawa, Hiroyuki Kitagawa

２．発表標題
Fast Algorithm for Integrating Clustering with Ranking on Heterogeneous Graphs

３．学会等名
the 20th International Conference on Information Integration and Web-based Applications and Services（国際学会）

４．発表年
2018年

１．発表者名
Hiroaki Shiokawa, Tomokatsu Takahashi, Hiroyuki Kitagawa

２．発表標題
ScaleSCAN: Scalable Density-based Graph Clustering

３．学会等名
the 29th International Conference on Database and Expert Systems Applications（国際学会）

４．発表年
2018年

１．発表者名
佐藤 朋紀，塩川 浩昭，北川 博之

２．発表標題
グラフの構造情報を用いたObjectRankの高速化

３．学会等名
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

４．発表年
2019年

１．発表者名
松下 朋弘，塩川 浩昭，北川 博之

２．発表標題
メッセージ集約に基づくAffinity Propagationの高速化

３．学会等名
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

４．発表年
2019年

１．発表者名
山崎 耕太郎，塩川 浩昭，北川 博之

２．発表標題
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

３．学会等名
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

４．発表年
2019年

１．発表者名
真次 彰平，塩川 浩昭，北川 博之

２．発表標題
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

３．学会等名
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

４．発表年
2019年

１．発表者名 真次 彰平，塩川 浩昭，北川 博之
２．発表標題 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
３．学会等名 情報処理学会 第81回全国大会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
オーストリア	Grifith University			