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

平成　３０　年　　６　月　１２　日現在

機関番号：　１２１０２
研究種目：　研究活動スタート支援
研究期間：　2016～2017
課題番号：　１６Ｈ０６６５０

研究課題名（和文）大規模グラフの頻出部分構造を利用した高速な分析アルゴリズムの開発

研究課題名（英文）Fast Graph Analysis using Frequent Subgraphs

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・助教

研究者番号：９０７７５２４８

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：本研究課題の目的は実世界の大規模グラフに対する高速な分析手法を開発することである．本研究期間を通じて，実世界のグラフが持つ頻出部分グラフ構造を捉えることで，数億ノード規模のグラフを高速に分析できる手法を提案した．具体的には(1)高速なグラフクラスタリング手法の開発および(2)高速なランダムウォーク分析手法の開発に取り組み，従来手法を数十から数百倍高速化するアルゴリズムを実現した．本研究成果を学術雑誌論文および国際会議論文として発表した．

研究成果の概要（英文）：The goal of this project is to develop efficient algorithms for large-scale graphs by capturing frequent subgraph structures. Specifically, in this project, we tackled to compute graphs with more than one hundred million nodes within short a few seconds. We proposed two algorithms in this project: (1) A structural graph clustering algorithm, and (2) an efficient random-walk analysis on graphs. Our experimental analysis on large-scale graphs showed that our algorithms achieved from 10 times to 100 times faster than the state-of-the-art algorithms.

研究分野：　データベース

キーワード：　グラフ　データベース　アルゴリズム

１．研究開始当初の背景

近年，数億ノード規模の大規模グラフが数多く登場している．例えば，ソーシャルネットワーク Facebook の規模は 2015 年時点で 15.5 億人を超えたと報告されている．この例に限らず，Web や生物学などの幅広い分野で大規模グラフが登場しており，その利活用に向けた大規模グラフ分析の高速化は近年の重要な研究課題となっている．

大規模グラフ分析高速化に関する研究は，2000 年代後半から世界的に活発に行われてきた．従来研究では，データベース分野を中心とする並列処理手法や応用数理分野による離散数学的手法が主流な高速化手法であった．しかしながら，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なのが現状である．このように，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．

２．研究の目的

本研究の目的は実世界の大規模グラフに対する高速な分析手法を開発することである．本研究期間を通じて，図 1 に示した様な，実世界のグラフが持つ頻出部分グラフ構造を捉えることで，数億ノード規模のグラフを高速に分析できる手法を提案する．本研究では，大規模グラフの高速な分析手法構築に向けて，まず基盤となる頻出部分グラフ構造検出手法を構築する．また，頻出部分グラフ構造を活用した大規模グラフの高速なクラスタ分析手法，および，高速なランダムウォーク分析手法を開発する．

３．研究の方法

本研究の目的は，実世界のグラフから頻出部分グラフ構造を捉え，この構造に基づく枝刈り手法を構築することで，数億ノード規模からなる大規模グラフを高速に分析する手法を開発する．まず，頻出部分グラフ構造を利用した高速なクラスタ分析手法の開発に取り組み，提案アプローチの有効性を検証する．その後，高速なランダムウォーク分析手法の開発に着手する．

４．研究成果

(1) グラフクラスタリングの高速化

構造的類似度に基づくグラフクラスタリング手法 SCAN は，グラフ中からクラスタやハブ，外れ値を高精度で検出できるため幅広く利用されている．例えば，図 1 のようなグラフ構造が与えられた時に，SCAN は密に接続したノード集合をクラスタとして抽出し，疎に接続したノード集合をハブまたは外れ値とみなすことができる．

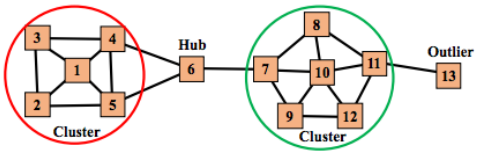

図 1．SCAN のクラスタリング結果

しかしながら，図 1 に示したようなクラスタリング結果を獲得するためには，SCAN は全てのエッジの構造的類似度と呼ばれる類似度指標を計算する必要がある．その結果として，計算量が$O(m^{1.5})$となり，大規模グラフの分析において膨大な計算時間を必要とする．

本研究では，SCAN においてボトルネックとなる構造的類似度の計算に着目し，この計算の中で処理時間を増大させる要素であるクリーク列挙処理の高速化を行った．具体的には，本研究では類似するクリーク構造に対する計算結果の枝刈りおよび，SIMD 命令を用いたデータ並列化を行う SCAN-XP を提案した．また，本研究では提案手法 SCAN-XP のさらなる高速化のため，メニーコアプロセッサ Intel Xeon Phi を活用した並列処理手法についても併せて検討を行った．

表 1 に示した数億ノード（数億エッジ）規模のグラフデータに対して，SCAN-XP ならびに SCAN の計算時間の評価を行った．図 2 に評価結果を示す．いずれの実行環境でも SCAN-XP が SCAN より高速であることがわかる．特に KNL 上で 272 スレッド実行された SCAN-XP は全てのデータセットにおいて SCAN より 100 倍以上高速であり，8 億 5 千万のエッジを持つ大規模グラフ webbase2001 を 36 秒で処理した．

表 1．データセット

Dataset	n	m
com-youtube	1,134,890	2,987,624
web-BerkStan	685,230	6,649,470
soc-Pokec	1,632,803	22,301,964
com-LiveJournal	3,997,962	34,681,189
soc-LiveJournal1	4,846,609	42,851,237
com-Orkut	3,072,441	117,185,083
webbase-2001	115,554,441	854,809,761

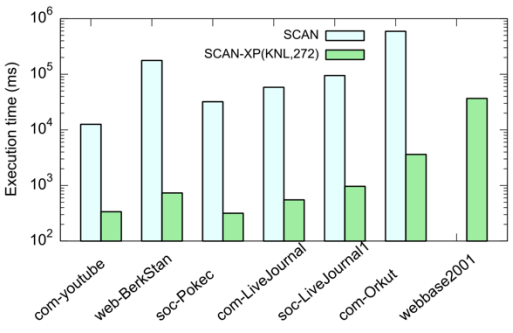

図 2．実行時間の比較

(2) ランダムウォーク分析の高速化
グラフ分析技術のひとつに，ObjectRank がある．ObjectRank は，ランダムウォーク分析手法 PageRank を拡張することでデータベース内のオブジェクトに対するキーワード検索を実現する手法である．データベース内のオブジェクトをグラフに見立てることによって PageRank と同様のリンク解析手法を適用し，キーワードに対する各オブジェクトの重要度を評価する．PageRank とは異なり，ObjectRank は複数の種類のノードとエッジからなるグラフを扱うことができるため，多様なデータに対して適用可能である．

ObjectRank は，クエリが与えられると行列ベクトル積の繰り返し演算によりグラフ全体のノードの重要度を評価する必要があるため，グラフ内のノード数を N，エッジ数を M，演算の繰り返し回数を T とすると，重要度評価の計算量は $O((N+M)T)$となる．ゆえに，ノード数が百万を超えるような大規模なグラフを対象としたとき現実的な時間でクエリ応答を行うことは難しくなる．

そこで本研究では，グラフに頻出するフリンジ構造において，ランダムウォーク分析におけるスコア値が早期に収束するという性質に着眼した．提案手法はこの性質に着目し，計算過程で ObjectRank スコアの収束値が閾値 ϵ を下回るノードを特定し，それらを逐次的にグラフから枝刈りすることで ObjectRank スコアの計算対象のノード数を削減する．枝刈り対象となるノードを効率的に発見するために，我々は理論的に ObjectRank スコアr_iの上限値$\overline{r_i}$と下限値$\underline{r_i}$を導出した．上限値$\overline{r_i}$ は $\overline{r_i} \geq r_i$の性質を満たすため，上限値が閾値 ϵ を下回ったとき枝刈り可能となる．また，下限値$\underline{r_i}$は上限値を$O(1)$で計算するために用いる．

図 3 に提案手法 FORank および既存手法 BinRank，ベースライン手法 ObjectRank の実行時間の比較結果を示す．実験結果より，閾値の値が大きい場合は枝刈りが上手く行われ，提案手法の方が約 10 倍程度高速となる．また，BinRank は提案手法と同様に高速であるが，BinRank は高速化を行うために 10 時間から 20 時間程度前処理を実行する必要がある．これに対して，提案手法 FORank は事前計算を必要としない．図 4 は FORank の各反復計算において計算されたノード数の割合を表している．この図からも分かるように頻出構造を活用した提案手法は効果的に計算対象エッジを削減できていることがわかる．

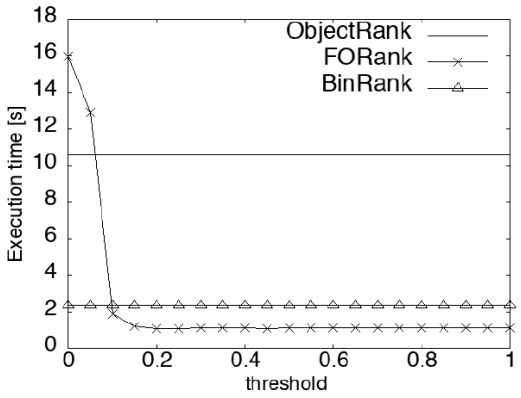

図 3. 実行時間の比較

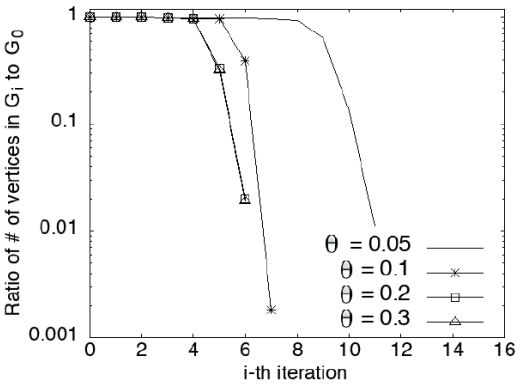

図 4. 計算されたノード数の割合

５．主な発表論文等
〔雑誌論文〕（計 6 件）全て査読有

[1] Tomoki Sato, Hiroaki Shiokawa, Yuto Yamaguchi, Hiroyuki Kitagawa,″FORank: Fast ObjectRank for Large Heterogeneous Graphs,″ In Proceedings of the 27th International World Wide Web Conference (WWW 2018), Lyon, France, April 2018. (印刷中)

[2] 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.

[3] 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.

[4] 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), pp.6:1-6:7, Chicago, IL, USA, May 2017.

[5] Makoto Onizuka, Toshimasa Fujimori, Hiroaki Shiokawa, "Graph Partitioning for Distributed Graph Processing," Data Science and Engineering, Vol.2, No.1, pp.94-108, February 2017.

[6] 藤森 俊匡, 塩川 浩昭, 鬼塚 真,"分散グラフ処理におけるグラフ分割," 情報処理学会論文誌：データベース (TOD72), Vol.9, No.4, pp.46-56, December 2016.

〔学会発表〕(計 11 件)
[1] 山﨑 耕太郎, 佐藤 朋紀, 塩川 浩昭, 北川 博之,"大規模グラフに対する逐次的なノード枝刈りを用いた RankClus の高速化," 情報処理学会 第80回全国大会, March 2018.

[2] 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 (DEIM2018), March 2018.

[3] 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.

[4] 松下 朋弘, 塩川 浩昭, 北川 博之,"セル分割による Affinity Propagation 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.

[5] 藤森 俊匡, 塩川 浩昭, 鬼塚 真,"分散グラフ処理におけるグラフ分割," 第 10 回 Web とデータベースに関するフォーラム (WebDB Forum 2017) , September 2017.

[6] 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10回 Web とデータベースに関するフォーラム (WebDB Forum 2017) , September 2017.

[7] 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," 第10回 Web とデータベースに関するフォーラム (WebDB Forum 2017) , September 2017.

[8] 藤森 俊匡, 塩川 浩昭, 鬼塚 真,"効率的な分散グラフ処理のためのグラフ分割," 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), March 2017.

[9] 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), March 2017.

[10] 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), March 2017.

[11] 佐藤 朋紀, 塩川 浩昭, 北川 博之,"大規模グラフに対する逐次的なノードの枝刈りを用いた ObjectRank の高速化," 情報処理学会 第79回全国大会, March 2017.

〔図書〕(計 1 件)
[1] Hiroaki Shiokawa, Makoto Onizuka, "Scalable Graph Clustering and its Applications," Encyclopedia of Social Network Analysis and Mining, 2nd edition, Springer, pp.1-10, New York, May 24th, 2017.

〔その他〕
ホームページ等
[研究代表者 Web ページ]
http://www.kde.cs.tsukuba.ac.jp/~shion/

[SCAN-XP ソースコード]
http://www.kde.cs.tsukuba.ac.jp/~shihakata/scanxp_codes/SCANXP_code.zip

６．研究組織
(1)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・助教
研究者番号：90775248