２０１７年度　博士研究員研究成果報告書

氏名（所属研究室）　照山　順一（理工学研究科加藤研究室）

研　究　課　題　Webグラフに対する定数時間アルゴリズムの開発と性能検証

研　究　期　間　２０１７年１０月１日～２０１８年３月３１日

研 究 成 果 概 要

インターネットを中心とする通信技術の発達により,10億を超える超巨大データが蓄積され,さらに膨張を続けている．近年のデータサイズの急速な増大はハードウェア性能の進歩を上回っており,従来のアルゴリズム設計では実用的な計算時間での情報解析が不可能となっている．そのため，このようなビッグデータを対象とした大規模問題群の解決には，アルゴリズム革新による解決が強く望まれている．アルゴリズム研究の分野において，多項式時間アルゴリズム（入力データのサイズをnとしたとき，nの多項式の時間で解を出す手法）というパラダイムが構築され，その下で多くの結果が出されてきた．従来ならば，多項式時間アルゴリズムであれば「速い」アルゴリズムであると考えられてきたが，現在のデータサイズに対してはn^2時間アルゴリズムでさえ，直接適用するだけでは実行時間やメモリ量など計算資源に関して大きな困難に直面する．つまり現実のデータに対してアルゴリズムが「速い」とは，計算時間が高々線形時間であることが求められている．

そのため,「劣線形時間アルゴリズム」という新しい計算パラダイムが提唱されている．その中でも,「定数時間アルゴリズム」は，入力データを定数だけの量を見るだけ，つまりほんの一部だけから問題を解こうという枠組みである．対象データがいかに巨大でも一定量のデータしか見ないため，ビッグデータを扱うのに理想的な手法であると考えられる．しかしながら，これまでの研究は理論的な興味に基づくものがほとんどであり，計算時間も定数とはいっても，天文学的に大きな定数であり，現実的な時間で計算が停止するとは到底いえないものが多く見られる．

定数時間アルゴリズムの存在が知られている問題でも，実際のネットワークの特徴を考慮しているわけではない．つまり，現実のネットワークが持つ特性を考慮することにより，天文学的な定数時間を実用的な定数時間まで改善することができるかもしれない．本研究では，Webページのリンクデータによるネットワーク（Web グラフ）に対し，ネットワークの特性を利用した定数時間アルゴリズムの開発及びその実装による実用可能性の検証を目指す．

定数時間アルゴリズムを実装するにあたり，重要となる理論的結果が２つある．Newman と Sohler によって超有限性を持つグラフを入力とする場合，いかなる性質でも定数時間で検査できることが示された．グラフが超有限性を持つとは，ε割合のリンクを切ることで大きさ定数の塊に分けられるこという.以下ではこのような分け方を「いい分割」と呼ぶこととする.Newmanらはランダムに選ばれた頂点が「いい分割」のどこに属するかを与える「分割オラクル」を構築し，任意の性質を検査する手法を与えた.「分割オラクル」の計算時間は定数時間ではあったが，天文学的に大きな定数となっており実装による実現性は低かった． Levi と Ron によって計算時間が効率的である「分割オラクル」が与えられ，本手法を用いることで現実のグラフデータに対して定数時間アルゴリズムを実装の実現性が高まったと考えられる．また，Levi らは「分割オラクル」設計に関して「いい分割」を得る全域分割アルゴリズムも与えている．

以上の結果から現実のグラフデータに対する定数時間アルゴリズム実装の実行可能性を検証

するためには，入力となるグラフデータが超有限性を持つか，つまり「いい分割」を持つかどうかの検証が不可欠である．報告者が所属する CREST 課題「ビッグデータ時代に向けた革新的アルゴリズム基盤」のメンバーである宇野裕之准教授研究室（大阪府立大学）のグループを中心に，Levi らの全域分割アルゴリズムの実装及びその拡張によって超有限性の検証が進められていた．研究期間以前における検証では，全域分割アルゴリズムの正確な実装を最重要課題とし，実行時間に関する考慮がなかった．そのため，頂点数が数万～数十万である中規模なグラフデータに対して計算時間を一日程度要した．しかし，全域分割アルゴリズムは乱数を用いたアルゴリズムであり，性能を正しく評価するには複数回の試行を行う必要があり，実行時間の高速化は重要な課題であり，報告者は実行時間の実用的改善に取り組んだ．

報告者は既存の実装を注意深く精査することでボトルネックとなっていた部分を見出し，適切なデータ構造を適用し分割アルゴリズムの実装を行った．この実装により，中規模なデータセットに対して数千倍の高速化が達成され，数秒～数十秒で計算可能となった．さらに，数百万頂点を持つ大規模なグラフデータに対しても検証を行うことが可能になった．現在，本実装を用いて多くのデータセットで超有限性の検証を推し進めている状況である．

また，既存の実装では Levi らの全域分割アルゴリズムを以下の３つの点で拡張している．（１）塊の大きさの上界を逐次的に増加，（２）大きさ上界を超える操作に対するペナルティの設定，（３）縮約対象となる最大重み枝の選択に乱数を使用．これら３点の拡張によって分割の性能は向上したものの，どの拡張が性能の向上に寄与しているのか明確になっていなかった．報告者は前述の高速プログラムの実装を拡張し，前述の拡張点を導入／非導入したプログラム８種の比較検証を進めている．頂点数数千～数十万の小規模・中規模グラフデータに関する実験は完了しており，その結果を検証したところ，（１）逐次的な上界の増加と（３）乱択選択は性能向上に寄与しており，（２）ペナルティは性能に寄与しないまたは悪化させることが分かった．

今後は拡張（１）（３）を導入した「分割オラクル」の実装が最大の課題である．「分割オラクル」の実装を基に，定数時間アルゴリズムの実装を目指し，超有限性を持つとみられるグラフデータに対する実験を推し進める予定である．