様式3

# 論 文 内 容 の 要 旨

| 氏 名 （ 新 井 淳 也 ） | |
|---|---|
| 論文題名 | クラウド環境における効率的なグラフクエリ処理に関する研究 |

論文内容の要旨

人々の活動や実世界のセンサから生成される大規模データ，即ちいわゆるビッグデータは多種多様な情報を含むことから，高い柔軟性を持つグラフ表現がしばしば用いられる．グラフの処理はデータ全体に対する不規則なランダムアクセスを生じさせるため，大規模なグラフに対するクエリを効率的に処理するためには大容量のメモリを搭載した計算機が必要とされる．そのような計算機の導入は高コストであったが，近年はパブリッククラウドのサービスが充実し大容量メモリの計算環境を安価に利用できるようになった．ますます大規模化していくグラフデータに対応するため，本研究ではクラウド環境におけるグラフクエリ処理に焦点を当てる．

クラウド環境の利用にあたってはグラフ中のパーソナルデータに関する流出リスクと法的な制約に配慮する必要がある．これらの問題を解消するため，本研究では次のようにk-匿名化を通じたクラウド利用の枠組みを考える．データ所有者は自身の計算環境においてパーソナルデータを匿名化し，それをクラウド環境へアップロードする．クラウド環境には高速なクエリ処理エンジンを配備し，匿名化されたデータに基づいてクエリを処理する．この枠組みを大規模なデータに対して適用するためには高速なクエリ処理エンジンと匿名化プログラムが必要である．しかしクエリ処理では不規則なメモリアクセスおよび処理結果に貢献しない無駄な探索が性能を低下させる．また匿名化においては，k-匿名性を得るために必要なデータの書き換え量を抑えつつ高速に処理することが困難である．

そこで本研究では高速なクエリ処理と匿名化を可能にするための3つのアルゴリズムを提案する．クエリ処理に対しては，メモリアクセスの局所性を向上させるためにグラフの頂点順序を最適化するアルゴリズム，および一度行った無駄な探索を二度繰り返さないよう枝刈りするアルゴリズムの2つを提案する．さらに匿名化に対しては，最近傍グラフの構築を通じデータ点のクラスタを発見することで書き換え量の削減と高速処理を両立したアルゴリズムを提案する．それぞれのアルゴリズムについて実験によってその有効性を検証すると共に，結果を踏まえ今後の研究課題についても議論する．

様式 7

# 論文審査の結果の要旨及び担当者

| 氏　名（　新井　淳也　） | | |
|---|---|---|
| | （職） | 氏　名 |
| 論文審査担当者 | | |
| 主 査 | 教授 | 鬼塚　真 |
| 副 査 | 教授 | 藤原　融 |
| 副 査 | 教授 | 原　隆浩 |
| 副 査 | 教授 | 松下　康之 |
| 副 査 | 教授 | 下條　真司 |

## 論文審査の結果の要旨

多様な情報に対応するため，あるいは表形式データとは異なる切り口でのデータ分析を実施するために，グラフ形式のデータが幅広く利用されている．ますます大規模化していくデータから迅速に情報を抽出するため，効率的なクエリ処理が必要とされる．グラフの処理は一般にグラフデータ全体に対するランダムアクセスを生じることから，メモリ上に全データを格納することが望ましい．大規模なグラフを一般的なサーバのメモリに格納することは困難であるが，近年はパブリッククラウドとして提供される大容量メモリを搭載した計算環境を安価に利用することができる．以上のような背景から，本論文はクラウド環境におけるグラフクエリ処理の効率化を目的としている．

クラウド環境におけるパーソナルデータの利用には情報漏洩のリスクや法規制が存在する．そこで本論文では匿名化したグラフデータをクラウド環境に配置し，そのデータに基づきクラウド環境でクエリを処理する枠組みを用いる．そのために必要となる効率的なクエリ処理と匿名化の実現にはそれぞれ課題が存在する．まずクエリ処理に関しては 2 つの課題がある．1 つ目は局所性の低さである．グラフ処理はデータ全体に対するランダムアクセスを伴うため，キャッシュミスが頻発し CPU の空転時間が長い．2 つ目は無駄な探索である．典型的なグラフクエリで用いられるサブグラフマッチングでは，しばしば検索条件に合致し得ない箇所に対してしらみつぶしの探索が行われ計算時間が増大する．さらに匿名化に関しては，低情報損失と高速処理の両立が課題となる．情報損失とは匿名性を得るためのデータの変動である．損失低減のためには全体最適な変動を加える必要があり，計算コストが高い．

本論文ではこれら 3 つの課題をそれぞれ解決する要素技術を提案している．主要な研究成果は次のとおりである．

1) 局所性を向上させるため，グラフデータに含まれる頂点の順序を最適化する高速なリオーダリングアルゴリズムが提案されている．既存のリオーダリング手法は局所性が高い頂点順序の生成に長い時間を要するのに対し，提案手法はグラフ中の階層的コミュニティ構造から効率的に局所性を抽出することで，高い局所性と短時間でのリオーダリングを同時に実現した．その結果，前処理としてのリオーダリングと本来目的としていたグラフ処理の合計時間という観点での高速化が可能になった．その観点における PageRank の平均 2.2 倍高速化，及びその他のグラフ処理 5 種の高速化が実験で確認されており，クエリ処理を含む全般的な高速化効果が見込まれる．
2) 無駄な探索を削減するため，探索失敗の履歴に基づいた枝刈りを行うサブグラフマッチングアルゴリズムが提案されている．既存手法はグラフ構造の事前解析によって枝刈り条件を導出するのに対し，提案手法は探索開始後にマッチ箇所を発見できなかった探索の過程から枝刈り条件を導出する．これにより枝刈りされる探索空間が劇的に増加し，既存手法と比べ最大 4 桁以上探索を高速化可能であることが実験で確認された．
3) 低情報損失と高速処理を両立するため，最近傍グラフを用いた$k$-匿名化のためのクラスタリングアルゴリズムが提案されている．$k$-匿名化の情報損失を低減するためには，似通った情報を持つ$k$人以上のグループを作成する必要がある．提案手法は最近傍グラフの構築を通じて人物のクラスタを検出することで情報損失の小さいグループを作成する．これにより既存手法比で最大 16%情報損失を減少させた．さらに多次元空間の分割による高速なグループ化手法を組み合わせることで，既存手法と同等の情報損失を約 1/10 の時間で達成した．

以上のように，本論文はグラフ処理に関する先駆的な研究として情報科学に寄与するところが大きい．よって本論文は博士（情報科学）の学位論文として価値のあるものと認める．