様式3

論 文 内 容 の 要 旨

氏 名 （ 伊藤 竜一 ）	
論文題名	複雑なスキーマを持つデータを管理するためのカーディナリティ推定に関する研究

論文内容の要旨

計算機資源の発展やデジタルトランスフォーメーションの進展に伴い，多種多様なデータが蓄積されその利活用が進んでいる．
これまで管理されなかった事象まで扱うことができるようになり，データの量だけでなく構造の複雑さも増している．
それらデータを利活用するトレンドと同時に，一個人からすると知られたくないようなデータ，つまりパーソナルデータが含まれることも増えていることから，プライバシ保護への配慮や流出・濫用を防ぐための規則の遵守も求められている．
これらの背景から，本研究では複雑なスキーマを持つ大規模データの効率的な処理とプライバシ保護に焦点を当てる．

まず複雑なスキーマを持つ大規模データの処理に対応するという観点では，高速化が必要となる．
これには大きく分けて2つのアプローチが考えられ，いずれかまたは両方利用できることが望ましい．
1つは既存のデータ処理技術を高速化するアプローチである．
ユーザから見ると変更を加えることなく高速になり，より大きなデータを扱えるようになるという点で有益である．
もう1つは処理結果の厳密性を捨てる代わりに前者以上の高速化をするアプローチである．
近似的な処理結果になるものの，厳密性が不要なユースケースであればより一層の高速化を期待できる．
次にパーソナルデータの処理という観点では，処理結果から個人の特定につながる情報が得られないようにする必要がある．
巧妙な攻撃手段が増え，攻撃者がどのように処理結果を悪用するのかや外部知識を利用する可能性などを事前に想定することは難しいため，特定の設定によらないプライバシ保護が望ましい．

本研究ではこれらの背景を踏まえ，複雑なスキーマを持つ大規模データ処理の高速化とプライバシ保護を実現するための3つの要素技術を提案する．
まず複雑なスキーマを持つ大規模データの高速な処理に繋がるカーディナリティ推定と呼ばれる技術に注目し，(1) 高速で精度の高いカーディナリティ推定手法，(2) 複雑なスキーマを持つ大規模データに対応するカーディナリティ推定手法に取り組む．
また，プライバシ保護として，厳密な指標である差分プライバシに注目し，(3) (1) や (2) に適用可能な差分プライベート学習手法に取り組む．

(1) では，データの分布を捉える深層学習技術に着目し，高速で精度の高いカーディナリティ推定手法を提案する．
カーディナリティ推定とはデータベース内に指定された条件を満たすデータが何件存在するかを推定するというタスクである．
カーディナリティ推定手法の改善により，既存のデータベースシステムはインターフェースを変えることなくより高速な処理手段を選択できるようになる．
また，カーディナリティ推定は条件に一致する件数を近似的に求める処理と等価であり近似的な処理のプリミティブとなることから，様々な近似的な処理の改善にもつながる．
従来のカーディナリティ推定手法は実世界データにそぐわない仮定を置いているため精度が低いことが知られている．
これを解決するため深層学習を利用した手法も提案されているが，人手によるパラメータ設定に依存している部分が大きく性能が安定していない．
提案手法では既存の深層学習を利用したカーディナリティ推定手法とは異なり，クエリに応じた推論を行うことで安定した推論を実現する．
実験では提案手法が安定して性能が高く高速に動作することを確認した．

(2) では，データベースシステムでカーディナリティ推定を利用する際に特に処理性能への影響が大きいと知られている，条件に結合を含むカーディナリティの推定を改善する提案を行う．
従来のカーディナリティ推定手法では，スキーマに存在する全てのテーブルに跨るデータの分布を，1つの推定器で捉えるか，テーブル間の相関に基づいた分割ごとの推定器で捉えることで結合を扱っている．

しかしながら，いずれのアプローチもスキーマの規模に対してスケールせず，推定精度の大幅な低下やそもそも推定が困難になるといった問題がある．
これに対し提案手法ではスキーマに定義された外部キー制約に基づいた分割ごとの推定器を用いることで，大規模かつ複雑なスキーマを持つデータに対応するカーディナリティ推定を実現する．
実験では既存手法が動作しないような大規模なスキーマを持つベンチマークで動作し，更に推定性能が高いことを確認した．
また，データベースシステムへの応用を想定したときに性能向上への寄与があることも確認した．

(3) では，(1) (2) で用いるような深層学習手法に対して，手法の有用性低下が少なく差分プライバシを満たすことができる手法を提案する．
差分プライバシとは，k-匿名化のように事前に攻撃パターンを想定するのではなく，識別困難性から，任意の背景知識を持つ攻撃者の任意の処理に対して安全性を担保する指標である．
従来の手法では差分プライバシを満たすことにより安全性を担保する一方で，深層学習手法自体の有用性，例えば分類器であれば分類性能が大きく低下してしまうという問題がある．
提案手法ではニューラルネットワーク内の大域的な冗長性と局所的な冗長性の両方を活用することで，安全性を下げることなく有用性の高い学習を実現する．
実験では，様々な深層学習モデル・タスクで差分プライバシを満たした有用性の高い学習となることを確認した．
また，(1) (2) のようなカーディナリティ推定手法に対しても有効であることを確認した．

様式 7

論文審査の結果の要旨及び担当者

氏　名　（　伊藤　竜一　）			
論文審査担当者		(職)	氏　名
	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

データの利活用が進み，データは大規模化が進んでいるだけでなく，より多様で複雑なデータ構造から構成されるものとなっている．大規模かつ複雑なデータから効率的に有益な情報を得るため，高速にクエリ処理する枠組みが必要とされている．更に，情報源となるデバイスのコモディティ化やストレージ容量の増加により様々なデータを蓄積できるようになったことで，扱うデータにパーソナルデータが含まれることも一般的になっている．一方パーソナルデータの扱いには法規制も進んでおり，適切なプライバシ保護が求められている．以上のような背景から，本論文は複雑なスキーマを持つ大規模データの効率的な処理とそのプライバシ保護を目的としている．

大規模かつ複雑なデータの効率的な処理のために，カーディナリティ推定の性能が課題となっている．既存のデータベースの場合，カーディナリティ推定の精度が低いと効率的なクエリ実行計画を作成することが困難となり，結果としてクエリ処理性能が低下してしまう．特に複雑なスキーマを持つデータでは結合処理が多数必要となるため，既存のカーディナリティ推定手法では推定精度が低下しやすい．カーディナリティ推定では，属性間の相関関係を如何に欠損なく捉えるか，結合クエリに対応するためテーブル間の関係を効率的に捉えられるか，また，それらの精度を保ちながらどうプライバシ保護を実現するかが課題である．本論文ではこれら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. Denoising Autoencoderを利用した高速で精度の高いカーディナリティ推定を提案する．提案手法では，データの分布をDenoising Autoencoderで捉えることで，属性間の相関関係を捉えながら適切な属性順での推論が可能となり，安定した高精度なカーディナリティ推定を実現している．複数のベンチマークにおいて推論速度と安定した推定精度の両立が確認されており，クエリオプティマイザや近似クエリ処理の改善が見込まれる．
2. 複数の密度推定器を組み合わせることで複雑なスキーマを持つデータに対して効率的な結合カーディナリティ推定を行う手法を提案する．提案手法では外部キー制約に基づいてスキーマを分割し，分割されたサブスキーマごとに密度推定器を学習する．推定時には複数の密度推定器を組み合わせた条件付き確率の推論を行うことで，テーブル数に対してスケーラブルな結合カーディナリティ推定を実現している．既存手法が大きな性能低下やメモリ不足で動作不良を起こすベンチマークでも推定性能が高く，クエリオプティマイザへの応用を想定した評価ではクエリ処理の高速化に寄与することが確認されている．
3. 深層学習で差分プライバシと呼ばれるプライバシ保護を効率的に満たす手法を提案することで，提案1，提案2で提案されているカーディナリティ推定をプライバシ保護化で動作可能にする．提案手法では，ニューラルネットワークのパラメータに存在する大域的な冗長性を低ランク近似で，局所的な冗長性をスパース化によって捉えて更新対象のパラメータを制限することで，プライバシ強度を落とすことなく有用性を高めている．カーディナリティ推定を含む深層学習モデルを利用した複数のタスクで既存の差分プライバシ手法と比較して有用性の向上が確認されている

以上のように，本論文はカーディナリティ推定とそのプライバシ保護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．