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　　元 年　　６ 月 ２３ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００１５４

研究課題名（和文）有用な仮説の自動探索・検証の高速化に関する研究

研究課題名（英文）Efficient framework for exploratory data mining

研究代表者
鬼塚　真（Onizuka, Makoto）

大阪大学・情報科学研究科・教授

研究者番号：６０７２６１６５

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：例外的なトレンドを探索するため，LOFを用いて例外的なOLAP分析結果を探索する方法を提案し，更にOLAP分析結果をグリッド分割することで，高速に例外OLAP分析結果を探索するアルゴリズムを開発した．提案手法では，信頼区間推定技術に基づいて各部分データの例外度の上限・下限を推定し，例外度上位 n 件に入り得ない部分データを探索処理の途中で足切りすることにより，効率的に局所例外部分データを特定する．評価実験の結果，提案手法は，既存の局所例外部分データ探索手法の探索時間を最大 84% 削減することに成功し，更にデータサイズに対するスケーラビリティを有していることを確認した．

研究成果の学術的意義や社会的意義
探索的データ分析の領域において，LOFを用いて例外的なOLAP分析結果を探索する方法は斬新なアイデアであり，且つ高速な探索を実現するためグリッド分割および信頼区間推定技術を組み合わせることで高スケーラビリティを達成した．論文誌1件採択，国際ワークショップ2件採択，表彰3件を受賞した（その他，国内シンポジウム5件），社会的意義としては，本技術を利用することで多様かつ大規模なデータに対して高速に有益な仮説の探索・検証が可能になるため，データサイエンティストを要せずビッグデータ解析が可能となる．現在，国立天文台と連携して超新星の発見応用に適用する準備を進めている段階であり，更なる発展が期待できる．

研究成果の概要（英文）：The goal of our research is to design a framework that effectively detect unexpected trends with regards to local outlier factor. The features of our framework is three-hold: 1) it is effective in detecting unexpected trends (expressed by OLAP queries) by leveraging local outlier factor, and 2) our framework introduces optimization techniques of top-k pruning and query sharing techniques over large number of answer candidates by employing online aggregation techniques for effective top-k pruning. Experiments results confirm that our method succeeds in reducing the search time up to 84% and also achieves high scalability with respect to input data size.

研究分野：データマイニング

キーワード：探索的データ分析　OLAP分析　データマイニング

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

1.　　研究背景

背景：近年，ビッグデータを分析することで隠れた知識やルールを発見して，社会的あるいは経済的なインパクトを生み出すことが期待されている．市場調査会社の IDC Japan の 2015 年 5 月の報告によれば，ビッグデータ分析に使われるインフラの国内市場が 2019 年までに 1469 億円に達する（年平均成長率 27%に相当）としている．その一方で，企業に対する現状のアンケート調査結果では，ビッグデータに関する技術を利用あるいは利用を検討している企業の数はこの 1 年で増加が収束傾向にあり，全体の 30%強という割合に留まっている．つまり，ビッグデータを分析できている企業はデータを分析する専門技術者であるデータサイエンティストを雇うことができるごく一部の企業だけであると推測することができる．また，総務省平成 26 年度版情報通信白書によればデータサイエンティストが不足していることが問題として指摘されている．このような背景を鑑み，本研究では分析領域の中でもビジネス業界で最も普及していることから社会インパクトが大きいと考えられる OLAP 分析の領域にターゲットを絞り，膨大な仮説集合を網羅的かつ自動的に探索・検証することでデータサイエンティストを要することなく OLAP 分析における仮説検証を可能にし，更に多様かつ大規模なデータに対する仮説の探索・検証を飛躍的に高速化する課題に挑戦する．

商品売り上げの分析の例を用いて，現状の OLAP 分析の問題を説明する．OLAP 分析では，地域毎あるいは月毎の観点において，商品全体および個々の商品の売り上げを分析するという分析が典型的分析方法である（group-by + aggregation クエリによる分析）．この場合，商品全体の売り上げ傾向と異なる特定の商品（全体データに対比して部分データと呼ぶ）を発見することが重要である．なぜならば，もし地域毎の傾向に関して，商品全体の売り上げと傾向が異なる商品を発見できれば，その商品は地域性の観点において標準的な売上げから乖離しており，地域性の影響が大きいと判断できるためである．このような判断結果に基づいて，ビジネス的な戦略として，該当の商品の調達量を地域毎に調整することで売上げを伸ばすことが出来る．例えば，図 1 にスイーツという特定の商品と全商品の売り上げを地域毎に比較した結果を示す．この結果から，スイーツの売上は全商品と比較して，千代田区では売れないが世田谷区・目黒区では多く売れていることが分かる．この結果に基づいて，例えば新商品のスイーツを世田谷区・目黒区で販売することを戦略的に決定することができる．しかし，現状の技術ではこのように全商品の売上げ傾向と異なる特定の商品を発見するには，分析者が試行錯誤的に商品を一つ一つ手動で選択して，商品毎にデータを分析し，その商品の分析結果がデータ全体の分析結果からどの程度乖離しているかを比較する必要があった．この問題を一般化して表現すると，試行錯誤的かつ手動で部分データを一つ一つ選択して分析を行い，全データに対する分析結果と比較する操作に膨大な時間を要することが従来技術の問題であり，これがビッグデータ利用の障壁となっていた．ここで部分データとは，任意の属性に対する任意の条件により選択される（例：商品カテゴリ＝スイーツ）ものを対象とする．このクラスの部分データ数は NU（但し，データの属性数を N，データの各属性値のユニーク値の平均件数を U とする）となるため，探索空間は膨大であり高速化が不可欠である．

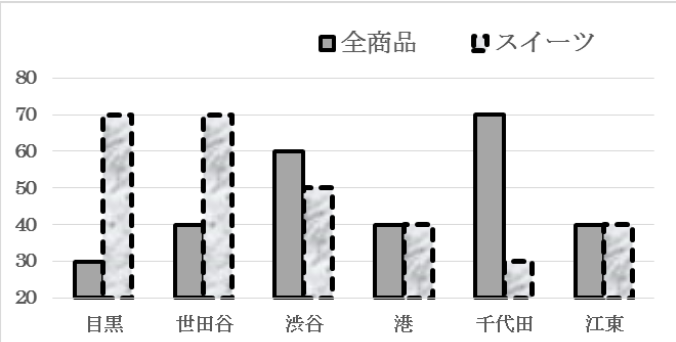

図1. コンビニにおける地域毎の売り上げ分析

２．研究の目的

ビッグデータによる分析が注目されているが，ビジネス業界で最も普及している OLAP 分析においては，データの多様化に伴い検証するべき仮説の件数が膨大になるため，有益な仮説を発見することは人手では困難かつ膨大な時間を要するという問題が生じている．本問題に対して，膨大な仮説集合を網羅的かつ自動的に探索・検証し，その処理を飛躍的に高速化するという課題に挑戦する．本課題を解決することで，人手による仮説の立案が不要になり，多様かつ大規模なデータに対して網羅的かつ高速に有益な仮説の探索・検証が可能になるため，データサイエンティストを要せずビッグデータ解析が可能となる．この結果，ビッグデータの活用が普及していない企業にビッグデータ分析を普及させることが可能となり，産業の発展に貢献することができる．

３．研究の方法

本研究課題では，データの多様化に伴って膨大になった仮説を網羅的かつ自動的に探索・検証する処理を飛躍的に高速化する課題に対し，２つの技術に関して研究に取り組む．

技術１：大量な仮説に関する探索・検証処理の共有化技術

仮説の探索を正確に行うという条件下において，全データに対する分析結果から最も乖離する分析結果を生み出す部分データを自動的かつ高速に探索・検証する技術を開発する．具体的

には，同時に複数の部分データに対する探索・検証をする処理を物理レベルにおいて共有化する，つまり分析対象のデータを１回シーケンシャルアクセスする際に，複数の部分データに対して全データに対する分析結果から最も乖離する分析結果を生み出すかどうかを判定することで，処理を高速化する．

技術２：信頼区間を活用した探索・検証処理のコスト削減技術

確率不等式を母集団平均の信頼区間の計算に適用することで，分析者が指定した信頼度を保証する条件下において，膨大な探索空間を大幅に削減して膨大な仮説集合に対する探索・検証を飛躍的に高速化する．更に，本研究成果の有用性を社会にアピールするため，経営科学系研究部会連合協議会が主催するデータ解析コンペティションに競技者として参画して，本研究成果の成功ユースケースを作り上げ社会にアピールする．

４．研究成果

[分析エンジン開発]

・膨大なクエリを探索する処理を高速化する複数クエリ共有化の技術と，有用性の高い結果を効率的に探索する top-k 検索の技術とを融合した探索的データ分析フレームワークを Spark 上に開発した．本フレームワークでは，データをストリーム処理して複数クエリを同時に処理するとともに，中心極限定理に基づいて有用性の低いデータキューブを推定して枝刈りすることを実現する．具体的には，データキューブを用いて統計情報を記録するとともに，データキューブと入力データの結合操作に対して hash join を用いて高速に統計情報を差分更新する機構を実装した．評価実験により，大域例外データを探索する場合，2 倍の高速化が可能であることを確認した．本成果は技術の高さが評価され国際ワークショップ DOLAP に採択された．加えて，本成果は 2016 年度のデータ解析コンペディション DB 部会にて最優秀賞を受賞，国内のシンポジウム DEIM2018 にて学生プレゼンテーション賞を受賞した．現在，論文誌に投稿を準備中である．

・有益な仮説を探索するため，LOF を用いて例外的な OLAP 分析結果を探索する方法を提案し，更に OLAP 分析結果をグリッド分割することで，高速に例外的な OLAP 分析結果を探索するアルゴリズムを開発した．提案手法では，信頼区間推定技術に基づいて各部分データの例外度の上限・下限を推定し，例外度上位 n 件に入り得ない部分データを探索処理の途中で足切りすることにより，不要なデータ読み込み量を削減することで効率的に局所例外部分データを特定する．評価実験の結果，提案手法は，既存の局所例外部分データ探索手法の探索時間を最大 84% 削減することに成功し，更にデータサイズに対するスケーラビリティを有していることを確認した．加えて本成果は，日本データベース学会論文誌に採択されており，国内のシンポジウム DEIM2018 にて学生プレゼンテーション賞を受賞した．

・データキューブを用いて大域例外探索と局所例外探索の中間結果を共有化することで，効率的に双方の探索を同時実行するフレームワークを完成した．本フレームワークは SparkSQL を用いて実装されており，高いスケーラビリティおよび近似探索の高い精度を達成した．

[実データ分析応用への適用]

経営科学系研究部会連合協議会が主催するデータ解析コンペティション（2016 年度および 2017 年度）に参画した．2016 年度は，実際にファッション EC サイトのデータに適用して有益な仮説が自動的に抽出できたことを検証した．本成果を DB 部会にて発表し，分析結果の良さが認められ最優秀賞を受賞した．2017 年度は，ヘアサロンの売り上げデータおよびファッション EC サイトの売り上げデータに対して，提案フレームワークを適用し有効性を検証した．その結果，地域性の観点および時間的観点それぞれにおいて例外的なふるまいをする分析データを自動的に探索できたことを確認した．本適用事例は，国際ワークショップ DARLI-AP に採択された．現在，国立天文台と連携して超新星の発見応用に適用する準備を進めている段階である．

〔受賞〕

1. DEIM2018 学生プレゼンテーション賞：論文タイトル：小笠原麻斗，松本拓海，佐々木勇和，鬼塚真，”統計的信頼区間を用いた局所例外部分データの効率的探索アルゴリズム”，データ工学と情報マネージメントに関するフォーラム(DEIM)，2018
2. DEIM2018 学生プレゼンテーション賞：論文タイトル：松本拓海，山室健，小笠原麻斗，佐々木勇和，鬼塚真，”探索的データ解析におけるエンジンの効率化”，データ工学と情報マネージメントに関するフォーラム(DEIM)，2018
3. データ解析コンペディション DB 部会 最優秀賞：論文タイトル：小笠原麻斗，松本拓海，水野陽平，佐々木勇和，鬼塚真，"局所例外部分データの自動探索，"ACM SIGMOD 日本支部，2017 年 2 月．

５．主な発表論文等

〔雑誌論文(査読有り)〕（計　1 件）

1. 小笠原麻斗，水野陽平，佐々木 勇和，鬼塚 真，局所例外部分データの自動探索，日本データベース学会和文論文誌，Vol.16，No.13，March 2018

〔国際ワークショップ(査読有り)〕（計　2 件）

1. Takumi Matsumoto, Yuya Sasaki, Makoto Onizuka, Data Slice Search for Local Outlier View Detection: A Case Study in Fashion EC, In Proceedings of International Workshop on Data Analytics Solutions for Real-Life Applications (DARLI-AP), March 2019
2. Yohei Mizuno, Yuya Sasaki, Makoto Onizuka, Efficient Data Slice Search for Exceptional View Detection, International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP), March 2017

〔学会発表(査読有り)〕（計　1 件）

1. 松本 拓海，小笠原麻斗，山室 健，佐々木 勇和，鬼塚 真，探索的データ分析におけるフレームワークの効率化，The Second Cross-Disciplinary Workshop on Computing Systems, Infrastructures, and Programming，2018

〔学会発表(査読なし)〕（計　4 件）

1. 松本 拓海，小笠原麻斗，山室 健，佐々木 勇和，鬼塚 真，大域的・局所的データ分析を両立した効率的なフレームワーク，データ工学と情報マネージメントに関するフォーラム(DEIM)，2019
2. 松本 拓海，山室 健，佐々木 勇和，鬼塚 真，探索的データ解析におけるエンジンの効率化，データ工学と情報マネージメントに関するフォーラム(DEIM)，2018
3. 小笠原麻斗，松本 拓海，佐々木 勇和，鬼塚 真，統計的信頼区間を用いた局所例外部分データの効率的探索アルゴリズム，データ工学と情報マネージメントに関するフォーラム，2018
4. 小笠原麻斗，水野陽平，佐々木 勇和，鬼塚 真，局所例外部分データの自動探索，データ工学と情報マネージメントに関するフォーラム(DEIM)，2017

〔その他〕
ホームページ等
https://github.com/OnizukaLab/exdatamining

６．研究組織
(1)研究協力者
研究協力者氏名：佐々木 勇和
ローマ字氏名：Yuya Sasaki

研究協力者氏名：山室 健
ローマ字氏名：Takeshi Yamamuro