情報処理学会論文誌 Vol.57 No.1 1–10 (Jan. 2016)

統計的信頼区間を用いた特徴的な部分データの効率的探索

水野 陽平[1,a)] 鬼塚 真[1,b)] 佐々木 勇和[1,c)]

受付日 2015年3月4日, 採録日 2015年8月1日

概要：ビジネスデータの解析において, 傾向や知識を発見するためには, OLAP（online analytical processing）型の集約・グループ化の分析処理を試行錯誤で繰り返し実行することが有効である．我々は特に，特徴的な分析クエリ結果を生み出す部分データを探索する分析に注目する（これを特徴部分データ特定問題と呼ぶ）．この問題を解くには探索範囲が膨大であるため，効率的な探索手法が不可欠である．そこで本稿では，1) 特徴部分データの自動探索手法と 2) 自動探索手法の最適化を提案する．最適化手法は，統計的技術の適用による探索範囲の削減と複数の SQL クエリの共有化による計算処理の削減を行っている．特徴部分データ特定問題の処理時間を計測した結果，最適化していない自動探索手法と比較して，最大 5 倍の高速化を確認した．

キーワード：探索的データ解析，確率論，OLAP

Exceptional Data Slice Search by Confidence Interval

Yohei Mizuno[1,a)] Makoto Onizuka[1,b)] Yuya Sasaki[1,c)]

Received: March 4, 2015, Accepted: August 1, 2015

Abstract: Business analysts repeatedly execute OLAP queries consisting of aggregate/group-by operations until they find trends, insights, and/or exceptions. One of interesting analysis is to identify particular data slices (subset of original data) that generate exceptional views apart from the average view generated from the whole original data. However, since the search space for identifying such data slices is quite huge, efficient techniques are indispensable for the data slice search. In this paper, we propose 1) a framework that automatically identifies data slices that generate outlier views for OLAP queries, and 2) an efficient algorithm that optimizes the data slice search. The algorithm reduces the search space by employing statistical techniques and removes redundant query processing by applying multi-query optimization for evaluating queries over multiple data slices. The experiments validate that our algorithm improves the performance five times faster than without the optimizations.

Keywords: Exploratory analysis, Probability theory, OLAP

1. はじめに

近年，企業が扱ってきたデータ以上に大規模かつ多様性に富んだデータであるビッグデータを資源として考え, 有益な情報を抽出する手法を適用することが重要な課題となっている．これまで OLAP（online analytical processing）やアソシエーションルールマイニング，クラスタリング，クラス分類，グラフマイニングなど多くのデータマイニング技術が開発されてきた [1]．特に OLAP 技術は，様々な次元からデータベースを集約・グループ化してデータの全体の傾向を把握したり，例外的なデータを発見することができるため，ビジネスデータの解析に頻繁に用いられる．ユーザは仮説・検証を繰り返すことにより，データから隠れた知識やルールを発見する．しかし，有用な仮説を生み出すことは，適切な分析対象属性と部分データを選択する必要があるため，データに関する深い理解が求められる．また，データの多様化に伴い検証するべき仮説の件数が膨

[1] 大阪大学大学院情報科学研究科
Graduate School for Information Science and Technology, Osaka University
a) mizuno.yohei@ist.osaka-u.ac.jp
b) onizuka@ist.osaka-u.ac.jp
c) sasaki@ist.osaka-u.ac.jp

大になるため，ユーザの大きな負担となっている．

探索的データ解析（Exploratory analysis）は上記の問題を解決し得るひとつの研究分野である [2], [3], [4], [5]．Morton ら [2] はユーザのデータ分析を支援するシステムについて調査している．Buoncristiano ら [3] は database exploration model に関する研究をしている．このモデルはユーザとシステムが対話することで，システムは標準的な分布から大きく乖離する分析クエリを特定することが大きな特徴である．Sarawagi ら [6][7][8] はデータマイニング技術を用いた OLAP キューブ分析に関する研究をしている．[6][8] は OLAP キューブから特異的なセルを探索する手法を提案している．しかし，これらは単一の値である OLAP キューブ内のセル（例えば 2014 年 3 月の衣服の売り上げ）を探索する手法であるため，乖離が大きい OLAP クエリ q（例えば月毎の売り上げ）を探索することができない．データを多次元的に分析することは，時期性や地域性など全体データの傾向を把握する上で重要である．また SEEDB[4], [5] は多次元データベースにおいて様々な OLAP クエリを探索し，全体データ D に OLAP クエリ q を実行した結果 $q(D)$ と部分データ S に実行した結果 $q(S)$ の乖離が最も大きくなるクエリ q を特定する．部分データは，属性がある特定の値といった条件で選択されるデータであり（性別 $=$′ 男性′ など），ユーザが指定する．しかし，SEEDB はユーザが指定した OLAP クエリ q を実行した際に $q(D)$ からの乖離が大きい $q(S)$ を生み出す部分データ S を探索するケースには適用することができない．このような部分データを探索する分析例として，全体の売り上げ平均から乖離している店舗（例えば若者に人気の原宿支店）や年度（例えばオリンピックイヤー）を特定する例などが挙げられる．しかし，部分データの探索範囲は膨大であるため（$O(|A| \times dv)$，$|A|$ は探索する属性数，dv はユニークな属性の取り得る値数である．），効率的な探索方法が不可欠である．

上記の部分データの探索範囲が膨大になる問題を解決するために，本稿では自動で特徴的な部分データ探索するフレームワークと最適化した効率的な探索方法を提案する．提案する探索方法の主な特徴は，統計的技術の適用による探索範囲の削減，および複数の部分データを探索する際に複数クエリの最適化技術を適用することによる無駄な計算処理の削減である．

1.1　分析例

売上データの地域性や時期性の影響を見て販売戦略を決める場合，分析処理に対して有用性が高い分析結果を生み出す特徴的な部分データを探索することが重要である．本稿ではこれを**特徴部分データ特定問題**と呼ぶ．具体例として，企業の販売データの分析を用いて，特徴部分データ特定問題の重要性を説明する．売上データを分析する場合，

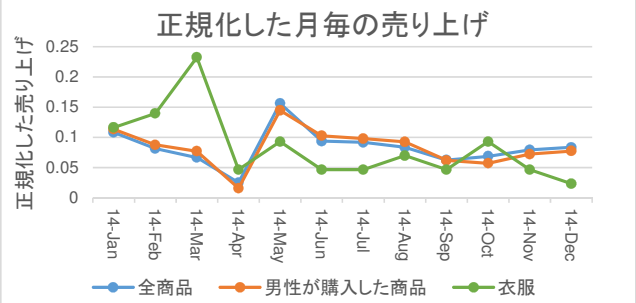


図 1: 全商品，男性が購入した商品，衣服の正規化した月毎の売り上げ

商品全体の売り上げの傾向からの乖離が大きい商品を見つけることが重要である．例えば商品全体の月毎の売り上げ傾向と異なる売り上げ傾向をもつ商品は時期の影響が大きいと判断でき，調達量を月毎に調整することや売り上げが低い月に新商品の投入や PR を行うなどの販売方針を戦略的に決定することができる．図 1 は全商品，男性が購入した商品，衣服に関する商品の月毎の売り上げの遷移の可視化結果である（1 年の売り上げの総和が 1 になるよう Y 軸の値を正規化している）．男性が購入した商品の月間総売り上げの遷移は全商品の月間総売り上げの遷移と乖離が小さいため特徴的な情報ではない一方，衣服に関する商品の月間総売り上げの遷移は全商品の月間総売り上げの遷移と乖離が大きいため，衣服の売り上げは販売時期の影響を大きく受けていると判断できる．そのため，売り上げが落ちている 4 月に新たな割引施策などを実施することにより売上を伸ばすことができる可能性がある．

本稿の構成は，以下の通りである．2 章にて問題定義について説明し，3 章において統計的技術について説明し，4 章にて提案手法の詳細を示し，5 章にて提案手法の評価と分析について説明する．6 章にて分析可視化結果について説明し，7 章において関連研究について述べ，8 章にて本稿をまとめ，今後の課題について論ずる．

2.　特徴部分データ特定問題の概要

本章では特徴部分データ特定問題について説明する．2.1 節で特徴部分データ特定問題の定義，2.2 節で問題を解く工程について説明する．

2.1　問題定義

特徴部分データ特定問題では関係データベース D を対象とする．以降，D を全体データと呼び，全体データ D の部分集合を部分データ S と呼ぶ．全体データ D はレコード $X_1, X_2, \cdots, X_{|D|}$（$|D|$ は全体データ D のレコード数）から構成される．レコード X_i（$1 \leq i \leq |D|$）は集約属性の集合（値段，重量など）とディメンション属性の集合 B（商品カテゴリ，地域など）から構成される．ディメンション属性の集合 B はディメンション属性 $b_1, b_2, \cdots, b_{|B|}$（$|B|$

は全体データ D のディメンション属性の数）から構成される．また，レコード X_i の集約属性 m の値を X_i^m とする．ディメンション属性 b_i（$1 \leq i \leq |B|$）が Y であるレコードから成る部分データ S は以下の形式で定義される．

$$S := \sigma_{b_i=Y}(D) \tag{1}$$

但し，σ は関係代数における選択演算である．特徴部分データ特定問題では部分データの集合 $\mathbb{S}$ と，単一の集約属性 m とグループ化属性 g から成る OLAP クエリ q を事前に設定する．具体的には，部分データ集合 $\mathbb{S}$ と OLAP クエリ q は以下の形式で定義される．

$$\mathbb{S} := \bigcup_{i=1}^{|B|} \{\sigma_{b_i=Y}(D) | Y \in values(b_i)\} \tag{2}$$

$$q(S) := {}_gG_{a=f(m)}(S) \tag{3}$$

但し，g は D のディメンション属性，$values(b_i)$ は属性 b_i の取り得るユニークな値の集合，f は属性 m に対する集約関数である．集約関数 f は件数計算（COUNT），平均値計算（AVG），総和計算（SUM）などがある．${}_gG_{a=f(m)}$ は各レコードをディメンション属性 g でグループ化し各グループごとに集約関数 f を m に適用する処理，a はその集約値である．つまり，${}_gG_{a=f(m)}(S)$ は部分データ S に集約・グループ化処理を実行する OLAP クエリであり，その結果はグループ化の値と集約値を組としたシーケンス型である．特徴部分データ特定問題は，分析処理を全体データに実行して得た分析結果に対して，部分データ集合 $\mathbb{S}$ のうち，乖離度が大きい分析結果を生み出す上位 k 件の部分データ S を求める問題と定義する．

定義 1 (特徴部分データ特定問題) 集約属性 m, グループ化属性 g，k を指定し，以下を満たす S を求める．

$$\underset{S \in \mathbb{S}}{\overset{k}{\operatorname{argmax}}}\ deviation(q(D), q(S)) \tag{4}$$

但し，$deviation$ は乖離度を数値化する関数であり，事前に与えられるものとし，本稿ではユークリッド距離にて乖離度を計算する．ユークリッド距離の場合は 2 つのシーケンスを入力として，ユークリッド空間におけるシーケンス間の距離を計算する関数である．乖離度を数値化する関数として，マンハッタン距離や dynamic time warping（DTW）[9] などを用いることも可能である．

2.2 具体例

特徴部分データ特定問題の具体例を説明する．

(1) $q(D)$ の計算：ユーザが事前に指定したクエリ q を全体データ D に対して実行する．例えば 1 章で説明した企業の販売データを分析する場合では，q が月毎の売り上げを計算する処理であり，q は SQL 言語により以下のように記述できる．

$SELECT$ 受注年月, SUM(販売金額)
$FROM$ 受注テーブル
$GROUP\ BY$ 受注年月;

(2) 部分データ集合 $\mathbb{S}$ の取得：全体データ D に対して部分データ S を選択するクエリを実行する．1 章で説明した例では，全商品から男性が購入した商品を選択する処理が該当する．部分データの数は 性別 $=$′ 男性′ と カテゴリ $=$′ 衣服′ の条件で選択される 2 つのみ示しているが，実際の分析では (探索する属性数) × (ユニークな属性の取り得る値数) となるため部分データの数は膨大となる．

(3) 式（1）を用いた部分データの乖離計算：様々な部分データ $S \in \mathbb{S}$ に対して OLAP クエリ q を実行し，$q(S)$ を計算する．1 章で説明した例では，男性が購入した商品に関する月毎の売り上げを計算する処理が該当する．次に，各部分データの分析結果 $q(S), S \in \mathbb{S}$ と全体データの分析結果 $q(D)$ に関して，関数 $deviation$ を用いて乖離度を数値化する．例えば（1）で計算した全商品に関する月毎の売り上げと例に挙げた男性が購入した商品に関する月毎の売り上げとのユークリッド距離を計算する処理である．最後に，関数 $deviation$ で計算した数値が上位 k 件の部分データの結果を可視化する．例えばユークリッド距離の値が大きい部分データの分析結果を折れ線グラフとして表示する処理である．

3. 確率不等式による区間推定

本章では部分データ特定問題を解く際の部分データの探索範囲削減に用いる信頼区間推定技術と区間推定に用いている確率不等式について説明する．3.1 節では Hoeffding の確率不等式 [10] について，3.2 節では Hoeffding の確率不等式を用いた区間推定について説明する．

3.1 Hoeffding の確率不等式

確率不等式は，母集団の確率変数に対して，母集団の確率分布を仮定することなく，期待値や分散などの限定的な情報だけに基づいて，母集団の確率変数の和あるいは平均に関する上限確率の上界を評価するものである．

Hoeffding の確率不等式は，確率変数の期待値や有限の定義域が判明しているときに，確率変数の上限確率の上界を与える有用な確率不等式として知られている．母集団 D の集約対象属性 m の確率分布が未知の場合に，非復元抽出した標本 $x = x_1, x_2, \cdots, x_{|x|}$（大きさ $|x|$ のレコードの集合）の集約対象属性 m の平均値が母集団 D の集約対象

属性 m の平均値を t 超える確率は以下の式で表せる．

$$Pr(t) = Pr[\overline{x} - \mu(D) \geq t] \tag{5}$$

但し，$\overline{x} = \frac{1}{|x|}\sum_{i=1}^{|x|} x_i^m$, $\mu(D) = \frac{1}{|D|}\sum_{i=1}^{|D|} X_i^m$，$t > 0$ である．この確率 $Pr(t)$ に対して，$t > 0$ において以下の式で上界を与えることができる．

$$Pr(t) \leq \exp[\frac{-2|x|t^2}{(1-f_x)(max-min)^2}] \tag{6}$$

但し，$f_x = \frac{|x|-1}{|D|}$（母集団に対する標本の割合），$min = \min\{X_i^m\}(i = 1, 2, \cdots, |D|)$，$max = \max\{X_i^m\}(i = 1, 2, \cdots, |D|)$ である．

3.2 Hoeffding の確率不等式を用いた区間推定

信頼区間は，母数がどのような数値の範囲にあるかを確率的に示す方法である．母数とは，確率変数の分布を特徴付ける数である．式（6）より平均値 $\mu(D)$ の信頼区間の式を導出する．区間を推定するため，$\overline{x}$ の値が $\mu(D)$ の値より大きくなる場合だけでなく，小さくなる場合も考慮する必要がある．$\overline{x}$ が $\mu(D)$ より t 以上大きくなる，もしくは小さくなる確率は式（5）の $\overline{x} - \mu(D)$ に絶対値を付けた以下の式になる．

$$Pr'(t) = Pr'[|\overline{x} - \mu(D)| \geq t] \tag{7}$$

式 (7) は，式 (5) の $\overline{x} - \mu(D)$ に絶対値を付けた式なので，確率 $Pr'(t)$ は，式（6）の確率 $Pr(t)$ の 2 倍となり，$t > 0$ において以下の式で上界 α を与えることができる．

$$Pr'(t) \leq 2 \cdot \exp[\frac{-2|x|t^2}{(1-f_x)(max-min)^2}] = \alpha \tag{8}$$

式 (7)，(8) より，母集団の平均値 $\mu(D)$ の $100 \cdot (1-\alpha)\%$ 信頼区間（α は信頼区間の有意水準）を以下の式で導出することができる．

$$\overline{x} - t < \mu(D) < \overline{x} + t \tag{9}$$

但し，

$$t = \sqrt{\frac{(1-f_x)(max-min)^2(\log 2 - \log \alpha)}{2|x|}} \tag{10}$$

次に，母集団 D の集約対象属性 m の総和 $\eta(D)$ を推定する．母集団 D の件数は $|D|$ なので推定する必要はなく，総和 $\eta(D)$ は平均値 $\mu(D)$ の確率不等式の各辺に件数 $|D|$ を乗算した以下の式で推定できる．

補題 1 $|D|(\overline{x} - t) < \eta(D) < |D|(\overline{x} + t)$
但し，$t = \sqrt{\frac{(1-f_x)(max-min)^2(\log 2 - \log \alpha)}{2|x|}}$

3.2.1 部分データの件数の区間推定

特徴部分データ特定問題では部分データの集約値を計算する．部分データの探索範囲を削減する際に部分データの集約値を推定する必要がある．そのため，適用する母集団 D の部分データ $\sigma(D) \subseteq D$ に関する確率不等式を考える．部分データに関する確率不等式を導出するには，まず部分データの件数を推定する必要がある．補題 1 を用いて部分データの件数を推定するため，標本の要素 x_i が $\sigma(D)$ に属するか否かを判定する関数 $exist$ を以下のように定義する．

$$exist(x_i) = \begin{cases} 1 & (x_i \in \sigma(D)) \\ 0 & (otherwise) \end{cases} \tag{11}$$

$exist$ 関数を用いて，標本に対する $\sigma(D)$ に属するデータの割合 $\overline{s}$ は次のように定義される．

$$\overline{s} = \frac{|\sigma(x)|}{|x|} = \frac{1}{|x|}\sum_{i=1}^{|x|} exist(x_i) \tag{12}$$

式（10）に関して，$exist$ 関数の定義より最小値 $min = 0$，最大値 $max = 1$ となるため，母集団に対する $\sigma(D)$ に属するデータの割合 $\frac{|\sigma(D)|}{|D|}$ は以下の式で推定できる．

$$\overline{s} - t' < \frac{|\sigma(D)|}{|D|} < \overline{s} + t' \tag{13}$$

$$t' = \sqrt{\frac{(1-f_x)(\log 2 - \log \alpha)}{2|x|}} \tag{14}$$

また，式（13）に母集団の大きさ $|D|$ を乗算すると，部分データの件数 $|\sigma(D)|$ を推定する以下の式が導出できる．

補題 2 $|D|(\overline{s} - t') < |\sigma(D)| < |D|(\overline{s} + t')$
但し，$t' = \sqrt{\frac{(1-f_x)(\log 2 - \log \alpha)}{2|x|}}$

3.2.2 部分データの平均値の区間推定

式（9），（10）に補題 2 を適用することで，部分データの集約対象属性 m の平均値 $\mu(\sigma(D))$ を推定する式を導出する．つまり，母集団の大きさを $|\sigma(D)|$，標本の大きさを $|\sigma(x)|$ と設定することで部分データの集約対象属性 m の平均値 $\mu(\sigma(D))$ を以下の式で推定できる．

$$\overline{\sigma(x)} - t'' < \mu(\sigma(D)) < \overline{\sigma(x)} + t'' \tag{15}$$

但し，$\overline{\sigma(x)}$ は標本中の $\sigma(D)$ に属するデータ $\sigma(x)$ の平均値，t'' は式（10）に補題 2 に関する以下の $|x|$ と f_x の条件を代入したものである．

$$|x| \leftarrow |\sigma(x)|$$

$$\frac{|\sigma(x)| - 1}{|D|(\overline{s} + t')} < f_x < \frac{|\sigma(x)| - 1}{|D|(\overline{s} - t')}$$

また，式（10）より t'' も区間 $[t_{min}, t_{max}]$ をもち，母集団の

大きさが $|D|(\overline{s}-t')$，$f_x=\frac{|\sigma(x)|-1}{|D|(\overline{s}-t')}$ のとき $t''=t_{min}$，母集団の大きさが $|D|(\overline{s}+t')$，$f_x=\frac{|\sigma(x)|-1}{|D|(\overline{s}+t')}$ のとき $t''=t_{max}$ となる．$t_{min}<t_{max}$ より式（15）は，以下の式で表現できる．

補題 3 $\overline{\sigma(x)}-t_{max}<\mu(\sigma(D))<\overline{\sigma(x)}+t_{max}$

但し，$t_{max}=\sqrt{\frac{(1-\frac{|\sigma(x)|-1}{|D|(\overline{s}+t')})(max-min)^2(\log 2-\log\alpha)}{2|\sigma(x)|}}$

3.2.3 部分データの総和の区間推定

部分データの集約対象属性 m の総和 $\eta(\sigma(D))$ の区間推定については，$\mu(\sigma(D))\cdot|\sigma(D)|$ を計算する必要がある．しかし，補題 2，3 より $\mu(\sigma(D))$ は $|\sigma(D)|$ に依存しているので，$\eta(\sigma(D))$ の最大値および最小値は必ずしも $\mu(\sigma(D))$ と $|\sigma(D)|$ の最大値同士および最小値同士の乗算とはならない．$|\sigma(D)|$ が最大値（$|D|(\overline{s}+t')$）をとるとき $t=t_{max}$ となり，$\mu(\sigma(D))$ も最大値（$\overline{\sigma(x)}+t_{max}$）をとる．つまり，$\eta(\sigma(D))$ の最大値は $\mu(\sigma(D))$ と $|\sigma(D)|$ の最大値同士の乗算である．しかし，$|\sigma(D)|$ が最小値（$|D|(\overline{s}-t')$）をとるとき $t=t_{min}$ となり，$\mu(\sigma(D))$ は最小値（$\overline{\sigma(x)}-t_{max}$）をとらない．そのため，$\eta(\sigma(D))$ が最小値となる $\overline{\sigma(x)}-t_{opt}$ と $|D|(\overline{s}-t'_{opt})$ を定義する．部分データの集約対象属性 m の総和 $\eta(\sigma(D))$ は以下の式で推定できる．

補題 4

$(\overline{\sigma(x)}-t_{opt})(|D|(\overline{s}-t'_{opt}))<\eta(\sigma(D))<(\overline{\sigma(x)}+t_{max})(|D|(\overline{s}+t'))$

但し，$t_{opt}=\sqrt{\frac{(1-\frac{|\sigma(x)|-1}{|D|(\overline{s}-t'_{opt})})(max-min)^2(\log 2-\log\alpha)}{2|\sigma(x)|}}$，

t'_{opt} は $\overline{s}-t'<t_{opt}<\overline{s}+t'$ の範囲で上記の $(\overline{\sigma(x)}-t_{opt})(|D|(\overline{s}-t'_{opt}))$ が最小となる値である．

4. 部分データ特定問題の高速化アルゴリズム

本章では，特徴部分データ特定問題に対して，1）統計的信頼区間推定の技術を用いることで，上位 k 件の部分データの探索範囲を削減する手法，2）複数クエリの共有化により無駄な計算処理を削減する手法の概要について説明する．

探索範囲削減手法

特徴部分データ特定問題における特徴として，大半の部分データの分析結果の傾向は，全体データの分析結果の傾向に類似していることが挙げられる．このような全体データの分析結果からの乖離が小さい部分データは，上位 k 件の候補になり得ないため，分析対象のデータの処理の途中において足きりすることが望ましい．そのため，統計的信頼区間推定の技術を適用して，上位 k 件の候補になり得ない部分データを早期に判断して，候補となる部分データの探索を足きりをすることで処理の高速化を図る．

具体的には，分析対象のデータの処理の途中において，統計的信頼区間推定の技術を適用することで分析結果の上限値と下限値を推定し，上位 k 件の候補になり得ない部分データの探索を足きりする．信頼区間の推定を行う方法としては，データを 1 パスでスキャンして実行する方法と 2 パスでスキャンして実行する方法が考えられる．1 パスの手法は全体データと部分データを同時に計算処理を実行しつつ部分データの足きりを行い，2 パスの手法は全体データの計算処理実行後，部分データの計算処理を実行しつつ部分データの足きりを行う．本論文では 2 パスの手法を用いて，信頼区間の推定を用いて足きりを行う方法を検討する [*1]．

探索範囲削減手法処理の流れは以下の通りである．

(1) 信頼区間の推定に用いるため，分析対象である入力データを標本とそれ以外のデータ集合に分割する．
(2) 標本に対して，部分データ毎に集約・グループ化処理を実行する．
(3) 実行結果に対して統計的信頼区間推定の技術を適用し，集約・グループ化結果の乖離度の上限値と下限値を推定し，上位 k 件の候補になり得ない部分データの足きりを行う．
(4) 標本以外のデータ集合に対して，上位 k 件の候補になり得る部分データの集約・グループ化処理を実行する．

複数クエリの共有化

特徴部分データ特定問題では，大量の部分データに対して集約・グループ化処理を実行するため，OLAP クエリの実行回数が膨大になる．同じ属性を対象としている場合，部分データの複数クエリを一つにまとめて実行することで高速化を図る．例えば，男性が購入した商品の月毎の売り上げと女性が購入した商品の月毎の売り上げを計算する処理では同じ性別という属性を用いて部分データ検索を行っている．このため，今回の例では 2 つのクエリを 1 つにまとめて実行することで無駄な処理を削減する．

4.1 探索範囲削減手法

本節では，Algorithm1 と補題 2, 3, 4 を用いて探索範囲削減手法の詳細について説明する．Algorithm1 は全体データの集約・グループ化処理結果 All，全体データの標本 F，グループ化属性 g，集約関数 f（SUM を仮定），集約対象属性 m，部分データを選択する条件に用いる属性 $b_1, b_2, \cdots, b_l$，有意水準 α，探索する部分データの件数 k を入力として，部分データの集約・グループ化結果 $Result$ を出力する．初めに，標本 F に対して，部分データ毎に集約・グループ化処理を実行する（1 行目）．次に，実行結果を用いて，部分データの集約・グループ化結果の区間推定を行う（5 行目から 8 行目）．事前に計算している全体データの集約・グループ化結果と推定した部分データの集約・グループ化結

[*1] 1 パスの方法は今後の課題とする

果を用いて，部分データ毎に乖離度の上限値と下限値を計算し（11 行目から 14 行目），上位 k 件の候補になり得ない部分データの足きりを行う（18 行目から 25 行目）．

Algorithm 1 信頼区間の技術を用いた部分データの足きり方法

Input $All, F, g, f, m, b_1, b_2, \cdots, b_l, \alpha, k$
Output $Result$

```
1: GroupByAggregate(Result, F, g, f, m)
2: for i = 1 to l do
3:   for each Y ∈ b_i do
4:     for each Z ∈ g do //各集約値の信頼区間を導出
5:       U_sum = U_ave U_num
6:       L_sum = L_ave L_num
7:       Result[Y][Z].UppBound ← U_sum
8:       Result[Y][Z].LowBound ← L_sum
9:     end for
10:    for each Z ∈ g do //乖離度の信頼区間を導出
11:      d_max ← Large(Result[Y][Z], All[Z])
12:      d_min ← Small(Result[Y][Z], All[Z])
13:      Score[Y].Best+= CalcuDistance(All[Z], d_max)
14:      Score[Y].Worst+= CalcuDistance(All[Z], d_min)
15:    end for
16:  end for
17: end for
18: threshold ← GetTopk(Score, k)
19: for i = 1 to l do //閾値以下の部分データの足きりを行う
20:   for each Y ∈ b_i do
21:     if Score[Y].Best < threshold then
22:       Result.Remove(Y)
23:     end if
24:   end for
25: end for
```

4.1.1 ユークリッド計算による乖離度計算

$q(D)$ と $q(S)$ の乖離度 $deviation(q(D), q(S))$ の信頼区間を導出する．式（3）の ${}_gG_{a=f(m)}(S)$ は以下の形式に展開できる．

$$ {}_gG_{a=f(m)}(S) := \{f(\pi_m(\sigma_{g='Z'}(S)))|Z \in values(g)\} \quad (16)$$

但し，$values(g)$ はグループ化属性 g の取り得るユニークな値の集合である．つまり ${}_gG_{a=f(m)}(S)$ は，*group-by* のグループ値 Z 毎に部分データを取り出し，その部分データの属性 m を射影して，集約関数 f を適用する処理である．定義 1 の $deviation(q(D), q(S))$ は式（16）を適用すると以下の形式に展開できる．

$$deviation(\{f(\pi_m(\sigma_{g='Z'}(D)))|Z \in values(g)\}, \{f(\pi_m(\sigma_{g='Z'}(S)))|Z \in values(g)\}) \quad (17)$$

また，$deviation$ はユークリッド距離を求める関数であるため，次元数 $|values(g)|$ の距離計算となり，式（17）は以下の形式で表現できる．

$$\sqrt{\sum_{Z \in values(g)} (f(\pi_m(\sigma_{g='Z'}(D))) - f(\pi_m(\sigma_{g='Z'}(S))))^2} \quad (18)$$

つまり，*group-by* のグループの値 Z 毎に D と S を選択操作して，属性 m に関して集約した結果の差を計算する．そして，得られた差を全グループの値全体で合計を算出する．

4.1.2 信頼区間の適用

式（18）に対して，2 パスの手法では 1 パス目において，$\{f(\pi_m(\sigma_{g='Z'}(D)))|Z \in values(g)\}$ の計算結果を記録し，$\{f(\pi_m(\sigma_{g='Z'}(S)))|Z \in values(g)\}$ に対して標本を利用して信頼区間の技術を適用する．f が $COUNT$，AVG，SUM の場合はそれぞれ補題 2, 3, 4 の区間推定を適用することができる．以下 f が $COUNT$ の場合を例として説明する．標本中の部分データに属するデータ集合を S とし，S を変数として補題 2 によって表現される区間 $range(|S|)$ を以下の形式で表現する．

$$range(|S|) = [|D|(\overline{s} - t'), |D|(\overline{s} + t')] \quad (19)$$

乖離度 $deviation(q(D), q(S))$ の信頼区間の上限値は $range$ を用いて以下の式で計算することができる．

$$\sqrt{\sum_{Z \in values(g)} (max(f(\pi_m(\sigma_{g='Z'}(D))), range(\pi_m(\sigma_{g='Z'}(S)))))^2} \quad (20)$$

但し，$max(x, [a, b])$ は値と範囲を引数とする関数であり，$max(|x - a|, |x - b|)$ を返却する．乖離度 $deviation(q(D), q(S))$ の信頼区間の下限値を計算する場合は，関数 $max(x, [a, b])$ の代わりに範囲 $[a, b]$ の中で x と最も距離が小さくなる値を返却する関数 $min(x, [a, b])$ を適用する．OLAP クエリ q の結果はシーケンス型であるため，$q(S)$ の各集約値の信頼区間を求める必要がある．標本 F と標本以外のデータ集合 R を以下のように表現する．

$$D = F \cup R \quad (21)$$

F は処理済みのデータ集合，R は未処理のデータ集合である．部分データ集合 $\mathbb{S}$ も同様に標本 $\mathbb{S}_F$ と標本以外ののデータ集合 $\mathbb{S}_R$ を以下のように表現する．

$$\mathbb{S} = \mathbb{S}_F \cup \mathbb{S}_R \quad (22)$$

$\mathbb{S}_F$，$\mathbb{S}_R$ はそれぞれ処理済みのデータ集合 F，未処理のデータ集合 R 内のデータで構成される部分データの集合である．補題 2, 3, 4 を用いて，${}_gG_{a=f(m)}(S_F)$ をもとに ${}_gG_{a=f(m)}(S_F \cup S_R)$ の信頼区間を推定する方法について説明する．f が $COUNT$（5, 6 行目の U_{count}, L_{count}），AVE（5, 6 行目の U_{ave}, L_{ave}），SUM（5, 6 行目の U_{sum}, L_{sum}）の場合について順に説明する．

(1) ${}_gG_{a=COUNT(m)}(S_F \cup S_R)$ を推定する場合：
$|\sigma_{g='Z'}(S_F \cup S_R)|$ を推定するために補題 2 を適用する．補題 2 の $|\sigma(D)|, \overline{s}, t'$ は，$|\sigma(D)| = |\sigma_{g='Z'}(S_F \cup S_R)|$，$\overline{s} = \frac{1}{|F|}\sum_{i=1}^{|F|} exist(F_i)$，$t' = \sqrt{\frac{(1-f_F)(\log 2 - \log \alpha)}{2|F|}}$ である．但し，$f_F = \frac{|F|-1}{|D|}$，$exist$ 関数は以下の式で表現できる．

$$exist(F_i) = \begin{cases} 1 & (F_i \in \sigma_{g='Z'}(S_F)) \\ 0 & (otherwise) \end{cases} \quad (23)$$

(2) ${}_gG_{a=AVE(m)}(S_F \cup S_R)$ を推定する場合：
$\mu(\sigma_{g='Z'}(S_F \cup S_R))$ を推定するために補題 3 を適用する．補題 3 の $\mu(\sigma(D)), \overline{\sigma(D)}, t_{max}$ は，$\mu(\sigma(D)) = \mu(\sigma_{g='Z'}(S_F \cup S_R))$, $\overline{\sigma(D)} = \overline{\sigma_{g='Z'}(S_F)}$, $t_{max} = \sqrt{\frac{(1-\frac{|\sigma_{g='Z'}(S_F)|-1}{|D|(\overline{\sigma}+t')})(max-min)^2(\log 2-\log\alpha)}{2|\sigma_{g='Z'}(S_F)|}}$ である．但し，$min = \min\{(S_F \cup S_R)_i^m\}(i = 1, 2, \cdots, |S_F \cup S_R|)$, $max = \max\{(S_F \cup S_R)_i^m\}(i = 1, 2, \cdots, |S_F \cup S_R|)$ である．

(3) ${}_gG_{a=SUM(m)}(S_F \cup S_R)$ を推定する場合：
$\eta(\sigma_{g='Z'}(S_F \cup S_R))$ を推定するために補題 4 を適用する．補題 4 の $\eta(\sigma(D)), t(t_{opt})$ は，$\eta(\sigma(D)) = \eta(\sigma_{g='Z'}(S_F \cup S_R))$, $t(t_{opt}) = \sqrt{\frac{(1-\frac{|\sigma_{g='Z'}(S_F)|-1}{|D|(\overline{\sigma}-t_{opt})})(max-min)^2(\log 2-\log\alpha)}{2|\sigma_{g='Z'}(S_F)|}}$ である．

乖離度 $deviation(q(D), q(S))$ の信頼区間を計算する．$q(D)$ と $q(S)$ を集約値の総和が 1 となるようにそれぞれ正規化する．乖離度の上限値を計算する際は，正規化した全体データの各集約値 $SUM(\sigma_{g='Z'}(D))$ と比較して，正規化した部分データの各集約値 $SUM(\sigma_{g='Z'}(S))$ の信頼区間中で距離が最も大きい値を採択する（11 行目）．正規化した $q(D)$ と採択した値で乖離度の上限値を計算する（13 行目）．同様に乖離度の下限値を計算する際は，正規化した全体データの各集約値 $SUM(\sigma_{g='Z'}(D))$ と比較して，正規化した部分データの各集約値 $SUM(\sigma_{g='Z'}(S))$ の信頼区間中で距離が最も小さい値を採択する（12 行目）．正規化した $q(D)$ と採択した値で乖離度の下限値を計算する（14 行目）．

4.1.3 部分データの足きり

特徴的な部分データ上位 k 件の候補になり得ない部分データの足きりを行う．標本の計算終了時に 4.1.2 項で導出した乖離度 $deviation(q(D), q(S))$ の信頼区間を計算する．全部分データにおいて上位から k 番目の乖離度の下限値（14 行目の $Score[Y].Worst$）を閾値として設定する（18 行目）．乖離度の上限値（13 行目の $Score[Y].Best$）が閾値を超えない部分データに関して足きりを行う（21, 22 行目）．

4.2 複数クエリの共有化

本節では，部分データの集約・グループ化処理結果を計算する複数の OLAP クエリを一つにまとめて実行することで無駄な計算処理を削減する方法を式（1）を用いて説明する．複数の OLAP クエリにおいて，部分データを選択する処理である $\sigma_{b_1=Y_1}(D)$ と $\sigma_{b_2=Y_2}(D)$ の選択する条件の属性が同じである場合（$b_1 = b_2$），1 つにまとめることができる．つまり，OLAP クエリ ${}_gG_{a=f(m)}(\sigma_{b_1=Y_1}(D))$ と ${}_gG_{a=f(m)}(\sigma_{b_1=Y_2}(D))$ は 1 つにまとめて（${}_gG_{a=f(m)}(\sigma_{b_1=\{Y_1,Y_2\}}(D))$）実行することができ，無駄な計算処理を削減することにより高速化を図る．

5. 評価実験

本章では提案手法の高速化の効果を評価する．5.1 節で実験方法について，5.2 節で実験結果について説明する．

5.1 実験方法

提案手法の有効性を検証するために，ナイーブな探索手法（NO_OPT），全ての OLAP クエリを 1 つにまとめて実行した場合（ONE_QUERY），複数クエリの共有化を適用した場合（MULTI），複数クエリの共有化と探索範囲削減手法を適用した場合（COMB）における実行時間を計測する実験を行った．乖離度を数値化する関数としてユークリッド距離を用いた．本実験のデータセットと用いた OLAP クエリは以下のとおりである．

データセット：経営科学系研究部会連合協議会主催，平成 27 年度データ解析コンペティションで提供されたデータである [*2]．レシート情報，レシート行番号，税込金額，点数，店舗，日付情報，時間帯，会員区分，性別区分など商品の受注に関する情報で構成されている．レコード数は 103,382,016（2 年），属性数は 32 である．

OLAP クエリ：集約属性は税込金額と点数，グループ化属性は店舗（9）と時間帯（19）を用いる．括弧の中の値は属性の取り得るユニークな値数である．

部分データの条件に用いる属性：部分データの条件に用いる属性は，店舗（9），時間帯（19），会員区分（2），性別区分（4），分類 1 コード（8），分類 2 コード（25），分類 3 コード（164）の 7 つとする．分類 1 コード，分類 2 コード，分類 3 コードは商品のカテゴリ区分の単位である．それぞれ分類 1 コードは分類 2 コードの，分類 2 コードは分類 3 コードの上位となっている．グループ化属性が店舗の OLAP クエリでは，時間帯，会員区分，性別区分，分類 1 コード，分類 2 コード，分類 3 コードを部分データの条件として設定した．グループ化属性が時間帯の OLAP クエリでは，店舗，会員区分，性別区分，分類 1 コード，分類 2 コード，分類 3 コードを部分データの条件として設定した．

本実験には，CPU が Intel(R) Core(TM) i7-4702MQ，クロック周波数は 2.20GHz, コア数は 4, メモリは 16GB の PC を使用し，データベース管理システムとして SQLServer2014 を用いた（非クラスター化カラムストアインデックスを適用）．また，データクリーニングのため集約属性の値が極端に大きいレコードは事前に取り除いた．

[*2] http://www.namalab.org/dac/data2_image.pdf

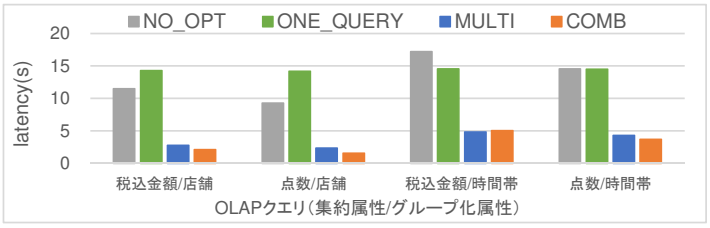


図 2: 4 つの手法の実行時間

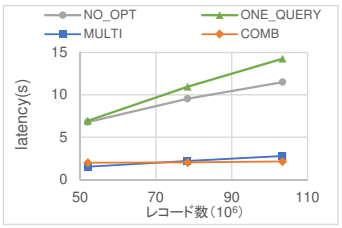


図 3: データサイズ変更時の実行時間

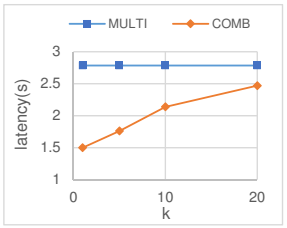


図 4: 上位 k 件変更時の実行時間

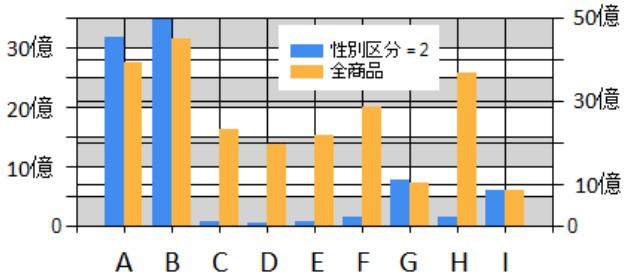


図 5: 店舗毎の売り上げにおける特徴的な部分データの分析結果 1（性別区分＝ 2）

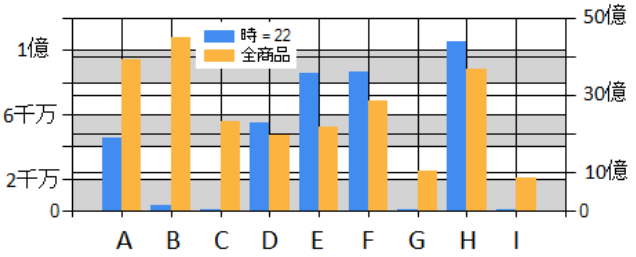


図 6: 店舗毎の売り上げにおける特徴的な部分データの分析結果 2（時＝ 22）

5.2 実験結果

図 2 は，様々な OLAP クエリを実行した際の特徴的な部分データ上位 10 件を探索する処理の実行時間である．X 軸は OLAP クエリ（集約属性/グループ化属性），Y 軸は実行時間（s）である．複数クエリの共有化，探索範囲削減手法ともに全ての OLAP クエリの実行時間削減に成功している．探索範囲削減手法はグループ化属性のユニークな値数が少ない際（店舗）に大幅な実行時間削減に成功している．これは，ユークリッド距離は次元数が大きくなるほど類似性が高くなる傾向があり，グループ化属性のユニークな値数が少なくなるほど足きりの機会が増えるためである．集約属性が税込金額，グループ化属性が時間帯の OLAP クエリの場合は，COMB は MULTI より高速化の効果が小さい．これは，上位 10 件の部分データの乖離度があまり大きくなく，多くの部分データを足きりできなかったためである．

図 3 はデータサイズ変更時の特徴的な部分データ上位 10 件を探索する処理の実行時間である．サンプリングサイズが大きくなると信頼区間の精度が向上するため，データサイズが大きくなるほど高速化の効果が期待できる．

図 4 は上位 k 件変更時の特徴的な部分データを探索する処理の実行時間である．k の値を小さくすると部分データの足きりの判定に用いる閾値が大きくなるので，より多くの部分データの足きりが可能となる．

全ての実験において，探索範囲削減手法を適用した場合としていない場合の上位 k 件の結果は一致している．

6. 分析可視化結果

本章では，提案手法をスーパーマーケットの販売データに適用して得られた分析結果について説明する．図 5,6 は店舗毎の売り上げにおいて特徴的な部分データの分析可視化結果の一例である．データセットは 2 年分である．右軸は全商品の売り上げ，左軸は部分データの売り上げの値である．図 5 の部分データは女性が購入した商品，図 6 の部分データは 22 時台に売れた商品，図 5,6 などの分析結果から他店舗と比べ，D 店，E 店，F 店，H 店は以下の共通の特徴を持つ．

- 女性の購入金額が少ない（他店舗に比べて男性の割合が大きい）
- 夜遅い時間帯の売り上げの割合が大きい
- 非会員の購入金額が多い

共通の特徴をもつ D 店，E 店，F 店，H 店の中で最も売り上げが少ない店舗は D 店，最も売り上げが多い店舗は H 店である．D 店の売り上げが少ない要因と H 店の売り上げが多い要因をより詳しく分析するため，それぞれ部分データの条件を D 店で売れた商品かつその他の部分データの条件，H 店で売れた商品かつその他の部分データの条件と設定し，月毎の売り上げにおいて特徴的な部分データを探索した．その分析可視化結果の一例が図 7,8 である．図 7 の部分データは D 店で売れた商品かつ 23 時台に売れた商品，図 8 の部分データは H 店で売れた商品かつ鮮魚に関する商品である．図 7 を見ると D 店は夜遅い時間帯の売り上げの割合が大きい店舗にもかかわらず，2014 年 11 月から急激に 23 時台の売り上げが落ちていることがわかる．この原因を分析し，改善することで似た特徴をもつ H 店の売り上げまでは D 店の売り上げを伸ばすことが期待できる．また，図 8 を見ると H 店の鮮魚に関する商品の売り上げは全商品の売り上げと異なる遷移を示しており，微増ながら 1 年目の売り上げと比較して 2 年目の売り上げが増加している．この傾向は D 店においてはあまり見られなかった．

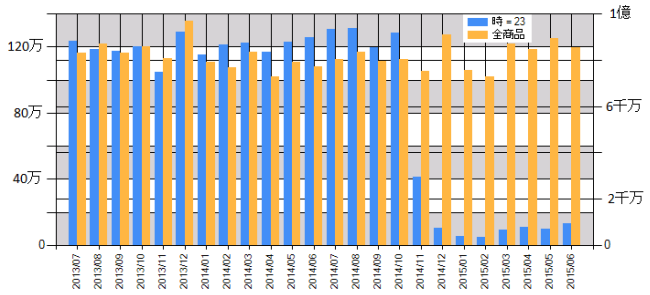


図 7: 月毎の売り上げにおける特徴的な部分データの分析結果 1（D 店，時＝ 23）

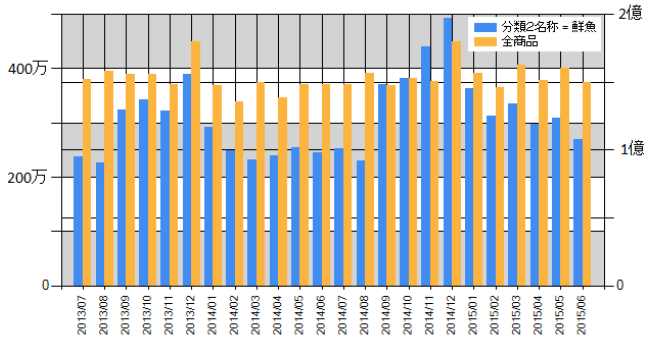


図 8: 月毎の売り上げにおける特徴的な部分データの分析結果 2（H 店，分類 2 名称＝鮮魚）

そのため，D 店において鮮魚に関する商品展開に力を入れることで D 店の売り上げを伸ばすことができる可能性がある．

7. 関連研究

データ解析ツールには，Spotfire[11]，Polaris[12] などがある．Spotfire は，散布図をベースとした可視化システムである．Polaris は基本的なデータベースクエリとテーブル代数による可視化の仕様を統合したシステムである．両者ともデータセットに最適な可視化設定を自動的に選択するが，分析者が設定することも可能である．これらのツールは分析者が着目したい全ての属性を手動で選択する必要がある．

自動で分析結果を可視化する機能を持つデータ解析ツールには，Profiler[13]，Vizdeck[14] などがある．Profiler はデータの異常を自動で検出し，いくつかの可視化結果を表示する．Vizdeck はダッシュボード上に 2 次元で表示し得る全ての可視化結果を表示する．

複数のソースからデータを収集・統合・可視化という一連の処理を自動化する技術として Google Fusion Tables[15]，DEVise[16] などがある．Google fusion tables は，web 上から様々なデータを収集し，統合することによりテーブルを作成し，その分析結果を可視化する．

データマイニングを用いて OLAP キューブを分析する研究が行われている [6][7][8]．Sarawagi ら [6][8] は OLAP キューブから特異的なセルを探索する手法を提案した．あるセルが特異的かどうかはキューブ内の他のセルから予測した値からの乖離によって定義される．これらの研究が OLAP キューブ内のセルを探索する一方，本研究では OLAP クエリの実行結果（シーケンス型）を探索する．

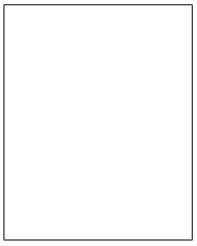
8. おわりに

本稿では，特徴的な部分データを効率的に探索する手法を提案した．提案手法では，まず問題設定を定義し，乖離度を数値化する関数としてユークリッド距離を用いることにより解く行程を説明した．さらに，統計的信頼区間推定の技術を用いて部分データの足きりを行い探索範囲を削減することと複数クエリの共有化に無駄な計算処理を省くことにより高速化を実現した．今後は有用性に関して更なる議論を行い，より詳細な分析のもと上位 k 件の分析可視化結果をユーザに提案する機能を実装する予定である．

参考文献

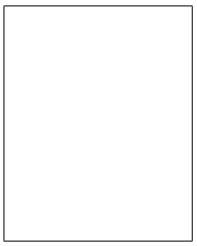
[1] Madraky, A.: Data Mining Text Book (2012).
[2] Morton, K., Balazinska, M., Grossman, D. and Mackinlay, J.: Support the data enthusiast: Challenges for next-generation data-analysis systems, *Proc. VLDB Endow.*, Vol. 7, No. 6, pp. 453–456 (2014).
[3] Buoncristiano, M., Mecca, G., Quintarelli, E., Roveri, M., Santoro, D. and Tanca, L.: Database Challenges for Exploratory Computing, *SIGMOD Rec.*, Vol. 44, No. 2, pp. 17–22 (online), DOI: 10.1145/2814710.2814714 (2015).
[4] Vartak, M., Madden, S., Parameswaran, A. and Polyzotis, N.: SeeDB: Automatically Generating Query Visualizations, *Proc. VLDB Endow.*, Vol. 7, No. 13, pp. 1581–1584 (online), DOI: 10.14778/2733004.2733035 (2014).
[5] Vartak, M., Rahman, S., Madden, S., Parameswaran, A. and Polyzotis, N.: SEEDB: efficient data-driven visualization recommendations to support visual analytics, *Proc. VLDB Endow.*, Vol. 8, No. 13, pp. 2182–2193 (2015).
[6] Sarawagi, S., Agrawal, R. and Megiddo, N.: Discovery-driven exploration of OLAP data cubes, *EDBT'98*, pp. 168–182 (1998).
[7] Sarawagi, S.: Explaining differences in multidimensional aggregates, *VLDB*, Citeseer, pp. 42–53 (1999).
[8] Sarawagi, S.: User-Adaptive Exploration of Multidimensional Data, *VLDB*, pp. 307–316 (2000).
[9] Müller, M.: *Information Retrieval for Music and Motion*, Springer-Verlag New York, Inc., Secaucus, NJ, USA (2007).
[10] Serfling, R. J.: Probability Inequalities for the Sum in Sampling without Replacement, *The Annals of Statistics*, Vol. 2, No. 1, pp. 39–48 (online), available from ⟨http://www.jstor.org/stable/2958379⟩ (1974).
[11] Ahlberg, C.: Spotfire: An Information Exploration Environment, *SIGMOD Rec.*, Vol. 25, No. 4, pp. 25–29 (online), DOI: 10.1145/245882.245893 (1996).
[12] Stolte, C., Tang, D. and Hanrahan, P.: Polaris: A System for Query, Analysis, and Visualization of Multidimensional Databases, *Commun. ACM*, Vol. 51, No. 11,

pp. 75–84 (online), DOI: 10.1145/1400214.1400234 (2008).
[13] Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M. and Heer, J.: Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment, *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, New York, NY, USA, ACM, pp. 547–554 (online), DOI: 10.1145/2254556.2254659 (2012).
[14] Key, A., Howe, B., Perry, D. and Aragon, C.: VizDeck: Self-organizing Dashboards for Visual Analytics, *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, New York, NY, USA, ACM, pp. 681–684 (online), DOI: 10.1145/2213836.2213931 (2012).
[15] Gonzalez, H., Halevy, A. Y., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., Shen, W. and Goldberg-Kidon, J.: Google Fusion Tables: Web-centered Data Management and Collaboration, *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, New York, NY, USA, ACM, pp. 1061–1066 (online), DOI: 10.1145/1807167.1807286 (2010).
[16] Livny, M., Ramakrishnan, R., Beyer, K., Chen, G., Donjerkovic, D., Lawande, S., Myllymaki, J. and Wenger, K.: DEVise: Integrated Querying and Visual Exploration of Large Datasets, *SIGMOD Rec.*, Vol. 26, No. 2, pp. 301–312 (online), DOI: 10.1145/253262.253335 (1997).



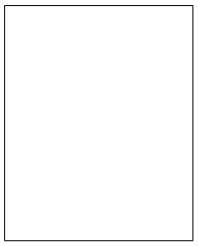
情報 太郎 （正会員）

1970 年生．1992 年情報処理大学理学部情報科学科卒業．1994 年同大学大学院修士課程修了．同年情報処理学会入社．オンライン出版の研究に従事．電子情報通信学会，IEEE，ACM 各会員．本会シニア会員．



処理 花子

1960 年生．1982 年情報処理大学理学部情報科学科卒業．1984 年同大学大学院修士課程修了．1987 年同博士課程修了．理学博士．1987 年情報処理大学助手．1992 年架空大学助教授．1997 年同大教授．オンライン出版の研究に従事．2010 年情報処理記念賞受賞．電子情報通信学会，IEEE，IEEE-CS，ACM 各会員．



学会 次郎 （名誉会員）

1950 年生．1974 年架空大学大学院修士課程修了．1987 年同博士課程修了．工学博士．1977 年架空大学助手．1992 年情報処理大学助教授．1987 年同大教授．2000 年から情報処理学会顧問．オンライン出版の研究に従事．2010 年情報処理記念賞受賞．情報処理学会理事．電子情報通信学会，IEEE，IEEE-CS，ACM 各会員．本会終身会員．