対話的データ分析におけるデータ分布の可視化手法について

平田 飛仙† 今村 誠† 菅野 幹人†

†三菱電機株式会社 〒247-8501 神奈川県鎌倉市大船 5-1-1

E-mail: †{Hirata.Takahisa@bp, Imamura.Makoto@bx, Kanno.Mikihito@bc}.MitsubishiElectric.co.jp

あらまし 対話的データ分析において，分析者が分析操作を行う際には，分析対象に関する知識が必要である．しかし，未知の大量データを分析する際には，分析者には分析操作を行うための事前知識が無く，またデータの全容を把握できないために，分析のきっかけが得られないという問題がある．筆者らは OLAP 分析を対象にこの問題について検討を行い，この問題が，データ抽出操作においてデータ抽出条件を決定するための知識を獲得する問題として現れることを明らかにした．また，この課題に対する解決策として，既存の OLAP 手法を拡張することで，絞り込み対象となるデータ分布を一括して算出し，俯瞰的に提示する可視化手法を提案し，実装した．

キーワード OLAP 分析，データ分布可視化，ユーザインタフェース，知識発見．

Data Distribution Visualization Method for Interactive Data Analysis

Takahisa HIRATA,† Makoto IMAMURA† and Mikihito KANNO†

†Mitsubishi Electric Corporation, 5-1-1, Ofuna, Kamakura, Kanagawa 247-8501, Japan

E-mail: †{Hirata.Takahisa@bp, Imamura.Makoto@bx, Kanno.Mikihito@bc}.MitsubishiElectric.co.jp

Abstract In this paper, we propose an data distribution visualization method for interactive data analysis. Interaction assumes users to have knowledge to manipulate, however if one try to analize unknown data, little knowledge is available. We study this problem in OLAP analysis, and pointed out that this problem emerges as a knowledge discovery problem to find suitable criteria to extract records. To solve this problem, we proposed a data distribution visualization method which presents distributions of all columns in OLAP cube. By contrast, ordinary OLAP analysis only presents a distribution of one target column. Finally, we present a data distribution visualization prototype which is an expansion of our OLAP tool with new features proposed in this paper.

Keywords OLAP analysis, data distribution visualization, user interface, knowledge discovery.

1 背景

1.1 対話的データ分析

近年，Visual Analytics などの呼称により，データの可視化とユーザインタフェースを融合した，対話的データ分析手法に関する研究が盛んとなっている[1]．対話的データ分析の特徴は，大量のデータをヒトが理解するための手段を提供することを目的としている点である．

対話的データ分析が注目されている背景としては，産業分野における情報化の浸透により，大量の業務データが蓄積される一方で，データ量がヒトに把握できる量を遥かに超えているために，蓄積されたデータを業務に活かすことができていないという問題が挙げられる．

従来の研究においては，データベース技術として，大量データの管理や高速な検索といった，分析の基盤となる機能について研究されてきた．また，データ分析技術として，特定の課題を対象にした様々な分析手法が研究されてきた．

しかし，実際の業務においてデータの分析を行う際には，大量のデータが蓄積されたデータベースから，特定の課題の分析に必要なデータを選別するという過程が必要となる．このようなデータベース技術とデータ分析技術の橋渡しを行うためには，解くべき課題に対して十分な知識を持った，ヒトによる媒介が不可欠となる．

対話的データ分析技術においては，このようなヒトによる媒介を支援することを目的としている点が，従来の研究と異なる点である．

1.2 産業分野における課題

産業分野においては，データ分析を行う場合，熟練者の勘やノウハウなどの，所謂「暗黙知」によって実現されている場合が多く見られる．これらの「暗黙知」は，分析対象の物理的特性の理解や，長年にわたる経験の蓄積などに基づいており，機械学習や統計手法などの汎用的な分析手法に勝る高い

精度を実現していることが普通である．

しかし，「暗黙知」による分析は，分析者の事前知識に強く依存しているため，新たな知見が得られにくいという課題がある．これは，「暗黙知」による分析では，分析対象となるデータを分析者の事前知識に基づいて選別するために，選別されたデータを分析しても，事前知識から想定される結果しか得られず，新たな知識が発見されにくいといったことが原因と考えられる．

この問題は，故障原因の分析や，改善策の検討といった，発見的な課題において特に問題となる．

一方で，近年では制御機器やセンサの電子化，コンピュータとネットワークの普及により，企業内には様々なデータが大量に蓄積されている．これらのデータの収集，維持，管理には多くのコストがかかるため，企業においては，大量に蓄積されているデータを有効活用することで，価値の創出に役立てたいという欲求がある．

しかし，これらのデータを新規な分析に活用したいと考えても，その量が膨大であるために，何処から分析をすべきかが分からずに手が付けられないという問題がある．

このため，これらのデータは，定型的な監視業務に用いられるか，上述のような「暗黙知」による限定的な分析に用いられる例を除けば，殆ど利用されておらず，ただ蓄積されているのが現状である．

このように，産業分野においては，既に大量のデータの蓄積があり，このようなデータを積極的に活用したいという欲求に加えて，活用すべき業務も存在するが，データが大量であるために，活用が進んでいないという課題がある．

1.3 本研究の狙い

著者らの目的は，大量に蓄積されたデータを，ヒトが利用可能にするための，対話的データ分析手法の開発である．

著者らは，これまでも対話的データ分析手法に関する開発を行ってきた[4][7]．現在，著者らは，対話的データ分析における課題として，分析対象データの選択や，データの可視化，分析操作の推薦などの手法について開発を行っている．

本論文では，対話的データ分析におけるデータの可視化に関する一案として，OLAP分析を対象に，分析対象データの絞り込み操作における分析者の支援を目的として，絞り込み対象のデータ分布を可視化する手法を提案する．

提案手法においては，前節で述べた課題における，分析対象のデータに関する事前知識の不足を，データ分布の可視化によって補うことで，未知の大量データに対しても，有効な分析を可能とすることを目指した．

2 問題提起

2.1 従来の OLAP 分析の課題

データベース分野においては，多数のカラムからなるデータベースを，多角的に分析するための手法として，OLAP (OnLine Analytical Processing) と呼ばれる概念が知られている[2]．OLAP分析においては，カラム毎に指定した抽出条件に合致するレコードを高速に抽出することで，対話的な分析手段を提供している．

しかし，OLAP 分析においては，抽出条件は，ユーザが任意に指定するものであり，その指定方法についてはあまり議論されてこなかった．そのため，OLAP 分析を行う際の課題として，どのように抽出条件を決めるかについては，分析者の事前知識に依存するという問題がある．

例えば，機器の寿命を OLAP 分析により分析することを考える．機器の寿命は機器が故障するまでの時間などにより定義され，寿命を決定する要因としては，機器の温度や負荷などが想定される．そこで，分析対象のカラムとしては，機器が故障するまでの時間，運転中の機器の平均温度，平均負荷率などが考えられる．

はじめに，機器の温度と機器が故障するまでの時間の関係を調べるべく，機器温度を横軸として，機器温度ごとに機器が故障するまでの時間を抽出し，平均故障時間間隔の分布を見ることを仮定する．このとき，機器温度の度数分布に関する知識が無ければ，どの程度の機器温度が「通常」であるのか，あるいは「高い」のか「低い」のかを判断することができないという課題がある．

つぎに，負荷率の影響を調べるべく，負荷率を抽出条件としてサンプルの抽出を行った後，同様に機器の温度と平均故障時間間隔の分布を見ることを仮定する．このとき，負荷率の度数分布に関する知識が無ければ，どの程度の負荷率が「通常」であるのか，あるいは「高い」のか「低い」のかを判断することができないために，負荷率の抽出条件を適切に決定することができないという課題が生じる．この様な状態で負荷率の抽出条件を決定するためには，抽出条件が非現実的な範囲にないか，抽出条件

が負荷率の分布に対して広過ぎたり狭過ぎたりしないかといった分布の状態を知るために，抽出条件を網羅的に走査する必要があり，煩雑な分析操作を要求されるばかりか，有意義な分析とならないといった問題が生じる．

同様の分析を，さらに多くのカラムを対象にして実施しようとすれば，これら複数のカラム間の相関により，1つのカラムの抽出条件を変更する度に，その他全てのカラムの分布が変化することになる．そのため，各カラムの抽出条件の妥当性を確認するためには，前記のような煩雑な操作を逐一実行する必要が生じるため，OLAP 分析は事実上実行不可能となる．

2.2 関連研究

OLAPの概念はCoddらによる定義が良く知られている[2]．Coddらによる定義では，直感的な操作や高速な処理などOLAP分析が備えるべき 12 の要件が定義されている．しかし，抽出条件を指定する際に提示すべき情報の要件については議論されていない．

OLAP分析における抽出条件の指定手法に関する研究としては，Leshらの研究や，Fujinoらの研究がある[3][4]．これらの研究では，抽出結果の分布に注目し，より特殊な分布形状が得られるような抽出操作を優先的に提示する手法について述べられている．これらの手法は，興味深い分布の得られる可能性が高い抽出条件を優先的に選択できる点が特長であるが，分析対象となるカラム内のデータ分布や，カラム間の関係などは分からないという課題がある．

また，絞り込み対象となるデータの可視化に関する研究としては，Tweedieらの研究がある[5][6]．これらの研究では，複数のカラムに対して，スライダーにより抽出条件を設定することで，各カラム内で抽出されたデータの分布を提示する手法について述べられている．これらの手法は，分析対象となるカラム内のデータ分布を可視化するという点で著者らの提案する手法と似ているが，OLAP分析のような自由度は無いという課題がある．

3 提案手法

3.1 データ抽出条件の探索

著者らは，OLAP 分析を以下のような手順に分けて考えた．

Step 1. 分析仮説の立案
Step 2. データ抽出条件の探索
Step 3. データ抽出条件の決定
Step 4. データ抽出の実行
Step 5. 分析仮説の検証

分析者は，Step1 において，分析意図に基づいて分析の仮説を立案し，Step2 において，この仮説を反映する抽出条件を探索し，Step3 において，Step2 で得られた知見を基に実際に抽出条件を指定する．Step4 では，Step3 で指定された抽出条件に従ってデータの抽出とプロット等が実施され，最後に，Step5 において，分析者は抽出結果を比較検討し，Step1 で立案した仮説の検証を行う．

前章の例では，「機器寿命は負荷率の大小に依存する」という仮説が，Step1 による分析仮説に相当し，どのような負荷率が「通常」であるか，あるいは「高い」か「低い」かなど，仮説検証に適した負荷率の抽出条件を探索する過程が Step2 に相当する．Step3 では，上記の探索により得られた知見により負荷率の抽出条件を決定し，Step4,5 にて，データを抽出し比較検討を行う．

従来の OLAP 分析においては，主に Step3,4 の機能について研究がなされてきた．一方で，Step2 の機能については，あまり注目されてこなかった．これは，データ抽出条件の探索が，Step1 において分析者の事前知識によって暗黙のうちに実行されるか，あるいは Step1 から Step5 の操作を通して獲得すべき知識と看做されているためであると思われる．

しかし，前章で述べたように，抽出条件を決定するためには，抽出対象となるカラム内のデータ分布やカラム間の相互の関係に関する知識が必要であるが，本研究が対象とするような未知の大量データの分析においては，分析者がそのような知識を事前に持つことは期待できない．また，抽出条件を決定するために，個々の抽出対象となるカラムに対して逐一網羅的な抽出条件の探索を行うとすれば，前章で述べたように，分析者の作業量が増大し，本来の分析目的を阻害する．

そこで，著者らは，データ抽出条件の探索を簡便に実施するための手法として，分析対象データの絞り込み操作において，絞り込み対象のデータ分布を可視化する手法を提案する．

3.2 実装方式

図 1は，提案方式で扱うデータ構造の模式図である．提案方式では，リレーショナルデータベースにおけるテーブル様のデータ構造を仮定している．

なお，OLAP 分析においては，高速にデータ抽出処

レコード →

カラム↓						
	分析対象カラム					
	絞り込み対象カラム 1					
	絞り込み対象カラム 2					
	絞り込み対象カラム 3					

図 1 データ構造の模式図

理を行うため，キューブと呼ばれる単位に分割して上記データを管理する．提案手法においても同様であるが，ここでは説明を割愛する．

従来の OLAP 分析においては，絞り込み対象カラムに対して抽出条件を指定し，抽出されたレコードから，分析対象カラムの集計結果を提示する機能が実現されている．

提案手法においては，上記に加えて，絞り込み操作を実行した際に，絞り込み対象カラムの集計結果を提示する機能を追加することで，絞り込み対象カラム内のデータ分布を可視化する方式とした．絞り込み対象カラムからデータを抽出する際の抽出条件としては，可視化対象となるカラム以外の絞り込み対象カラムに対して指定された抽出条件を用いた．例えば，絞り込み対象カラム 1 のデータ分布を可視化する際には，絞り込み対象カラム 2 と 3 に指定された抽出条件にてレコードを抽出し，絞り込み対象カラム 1 の集計結果を提示するといった具合である．

以上の説明から明らかなように，絞り込み対象カラムの可視化は，従来の OLAP 分析の機能によって実現されており，従来と異なる点は，これを事前に一括して計算し，分析者に提示する点である．

このような構成により，分析者は，絞り込み対象カラム内のデータ分布を俯瞰的に把握可能なため，カラム毎に個別に抽出条件の探索を行う必要が無く，分析効率の向上が期待される．また，ある絞り込み対象カラムに対して，抽出条件を変化させた際に，他の絞り込み対象カラム内のデータ分布が変化する様子が可視化されるため，絞り込み対象カラム間の関係が把握可能となることも期待される．

また，絞り込み対象カラムに対するレコードの抽出処理においても，各カラムに対する抽出結果が再利用可能なため，個別に計算を行う場合に比べて，少ない計算量で抽出処理が可能である．

3.3 プロトタイプ

今回開発したプロトタイプの基本機能は，従来の OLAP手法[3]を継承しており，高速なサンプル抽出

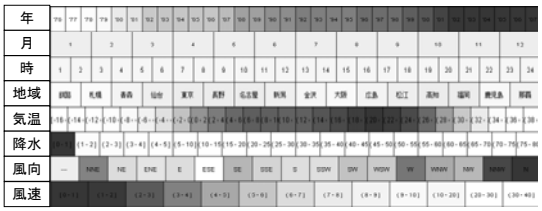

図 2 提案手法のプロトタイプ画面

性能とクリックベースの簡易な分析インタフェースを備えたツールとなっている．

今回は，提案方式の実装にあたり，前章までに述べた，以下の機能を追加実装した．

- 絞り込み対象データの抽出機能
- 絞り込み対象データ分布の可視化機能

図 2 は，提案方式を実装したプロトタイプの画面の一部である．本プロトタイプは，従来から我々が研究を行ってきたOLAP手法[3]を基盤として開発されており，図 2 は，従来手法においてサンプルの絞り込み条件設定を行うインタフェース部に提案手法を適用した例である．図において，色の濃淡が，サンプル数に対応しており，色が濃いほどサンプル密度が高いことを表している．

現在は，方式検証の段階のため，可視化機能としては，既存抽出条件指定のためのボタンに濃淡を付けるだけの実装である．今後は，グラフ表示による，より詳細な分布の提示機能や，絞り込みデータ分布に合わせた抽出条件指定範囲の動的な変更機能などの実装を検討している．

なお，この例は，気象庁が公開している全国の AMeDAS データ 30 年分を可視化した例となっており，年度が進むに従ってサンプル数が増えている様子や気温の分布がおよそ 20℃を頂点として高温側に偏りのある単峰性の分布をしいることなどが確認できる．

図 3 は，図 2 と同様の AMeDAS データを用いて簡単な分析を行った例である．図 3 においては，東京における，7 月，8 月，9 月の 3ヶ月間を対象に，気温が高い場合と低い場合の各気象要素におけるサンプルの分布を可視化した．気温の分布に注目すると，それぞれの場合において，分布の上端と下端を抽出対象としていることから，直ちに気温が「高い」場合と「低い」場合が選択されていることが理解できる．また，それぞれのサンプルの分布を比較することで，日中は気温が高く，明け方に気温が低いことが分かる．さらに，気温が高い時には南寄りの風

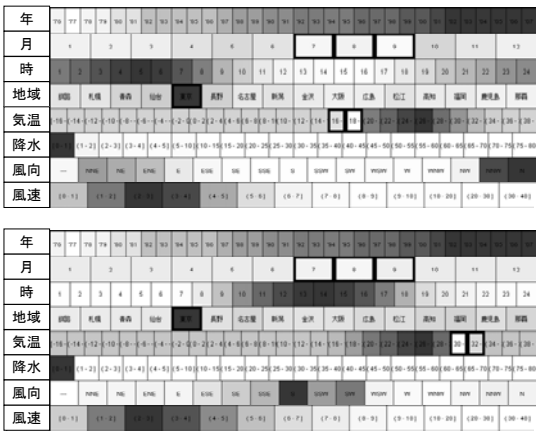

図 3　提案手法による分析例
[上:夏の東京における気温が低い場合の分布
下:夏の東京における気温が高い場合の分布]

が吹く傾向にあり，気温が低い時には北寄りの風が吹く傾向にあることが分かる．時刻と風向の関係を考えれば，これが海風と陸風に対応していることなどが推察される．

このような分析は，東京に限らず世界のいかなる都市に対しても即座に実行可能であり，提案手法を用いることで，地域性などの事前知識なしにデータ抽出条件の探索が容易に実施可能であることが分かる．同様に，産業分野においては，機器の違いや使用環境の違いなどの事前知識なしに絞り込み条件の探索が可能となるといった効果が期待される．

4　まとめ

本論文においては，対話的データ分析において，データ分布を可視化する手法について述べた．

とくに，OLAP 分析を対象に検討し，未知の大量データを分析する際には，抽出条件の探索を支援する機能の実現が本質的な課題であることを指摘した．

これに対し，著者らは，OLAP 分析において，分析対象データの抽出条件を指定する際に，絞り込み対象データの分布を可視化する手法を提案した．

また，提案手法の実装方式として，従来の OLAP 機能を拡張することで，絞り込み対象データの分布を効率的に可視化可能であることを述べた．

提案手法の特徴は，絞り込み対象データの分布を一括して計算し，提示することである．これにより，分析者は，絞り込み対象データの分布を俯瞰的に把握可能になるとともに，抽出条件を変化させた際の，絞り込み対象データ間の関係も把握可能となることが期待される．

現在は，提案方式を実装したプロトタイプを開発中であり，基本機能を実装した方式検証を完了している．

今後は，分布の提示方式の改善や分布に合わせた抽出条件指定の動的変更機能などの実装により，ユーザインタフェース機能を充実化し，フィールド試験による有用性の評価を実施予定である．

参　考　文　献

[1] D. Keim et al., “Mastering the Information Age: Solving Problems with Visual Analytics,” EuroGraphics Association, 2010.

[2] E.F. Codd, S.B. Codd and C.T. Salley, “Providing OLAP to User-Analysts: An IT Mandate,” Codd & Date Inc., 1993.

[3] N. Lesh and M. Mitzenmacher, “Interactive data summarization: An example application,” Proc. of Conf. on Advanced Visual Interfaces, ACM, 2004.

[4] T. Fujino et al., “Data Analysis System using Trend Visualization,” Proc. IEICE General Conference, 2005.

[5] L. Tweedie, R. Spence, H. Dawkes and H. Su, “Externalizing Abstract Mathematical Models,” Proc. of Conf. on Computer-Human Interaction, ACM, 1996.

[6] R. Spence and L. Tweedie, “The Attribute Explorer: Information synthesis via exploration,” Journal of Human-Computer Interaction, British Computer Society, 1998.

[7] T. Hirata et al., “Error Visualization Method for High-Dimensional Models,” Proc. of Forum on Data Engineering and Information Management, DEIM, 2010.