解説

知識創発につながるデータ分析環境†

砂山 渡*1・高間 康史*2

1. はじめに

昨今、データを有効に活用できるスキルを持った人材の育成が重視されており、平成 29 年度に文部科学省「成長分野を支える情報技術人材の育成拠点の形成（enPiT: Education Network for Practical Information Technologies）」、平成 30 年度に文部科学省「超スマート社会の実現に向けたデータサイエンティスト育成事業」などのプログラムも開始され、社会人の学び直し（リカレント教育）を推奨する動きがある。また大学においては、企業における情報戦略室に相当する、IR:Institutional Research に関する部署を設け、大学の経営改善や学生支援、教育の質向上のため、学内データを収集・分析し、改善施策を立案、施策の実行・検証を行う動きが見られる。

一方で、データ分析に不慣れな人は、専門的な知識がないとデータ分析を行うことは難しいと考えていたり、データ分析のためのツールを導入すれば、ツールが分析してすべての結果を出してくれるという誤解を抱いているケースが多い。そこで本稿では、データ分析とは何をする事なのか、どうすればデータ分析ができるようになるのかを最も基本的なところから述べるとともに、知識創発につながるデータ分析環境について概説する。

以下、本稿では、2 章でデータ分析による意思決定プロセスについて、3 章で知識に繋がれるデータについて述べる。4 章でデータ分析手法とその出力の可視化について、5 章で分析結果に意味を与える解釈のための可視化手法について述べる。6 章で解釈を集めるためのインタラクションの方法について、7 章で集めた解釈を知識につなげるための方法について述べ

図1 データ分析による意思決定支援のプロセス

る。8 章で知識創発プロセス支援の具体例について述べ、9 章で本稿を締めくくる。

2. データ分析による意思決定プロセス

本章では、データ分析による意思決定プロセスについて述べる。

2.1 データ分析による意思決定プロセスの概要

データ分析プロセスは、データマイニングという言葉が使われ始めた 1996 年頃に Fayyad らが、データを入力した後「データ選択」「前処理」「データ変換」「データマイニング」「解釈・評価」という処理の後に知識に至る流れを示している [1]。本稿においては、このプロセスをより精緻化したデータ分析による意思決定プロセス（図 1）をもとに話を進める¹。

本プロセスは、意思決定の判断根拠となる知識を得るまでの 11 の項目からなり、前半の 0. から 7. はデータを分析する「データ分析プロセス」、後半の 8. から 10. はデータを解釈する「知識創発プロセス」となる。以降の各節においては、各項目の詳細について述べて

† Data Analysis Environments for Knowledge Emergence
Wataru SUNAYAMA and Yasufumi TAKAMA

*1 滋賀県立大学工学部
School of Engineering, The University of Shiga Prefecture

*2 首都大学東京システムデザイン学部
Faculty of System Design, Tokyo Metropolitan University

1 ただし、「前処理」と「データ変換」は、本稿では「データ整形」としてまとめられている。

いく。

図1においては、コンピュータが自動的に行ってくれない、人間の知的思考が必要となる項目を下線で示している。全部で11の項目のうち、最初のデータが自動的に蓄積される項目0.と、入力データをツールが処理して可視化する項目6.7.以外は、人間の思考と作業が必要となる。しかし、データ分析に不慣れな人は、このプロセスの全体像を把握しておらず、中央の「6. データ処理」と、「7. データ可視化」がすべてであり、ツールを導入すれば、手元の「0. 蓄積データ」を入力することで、即座に知識が得られると誤解することがある。

しかし実際には、残りの8つの項目を実現する人間の知的思考が必要であり、項目によっては、データサイエンティストではなく、データから得られる知識を活用したい現場の人による思考が必要となる。データサイエンティストが、どれだけ分析の手法を熟知していたとしても、データが対象とするドメインの専門知識までを有しているケースは多くない。そのため、実際にデータが有効に活用される分析を実現するためには、1) ドメインの専門知識を有する人が、データ分析のスキルを身につける、2) データサイエンティストが、ドメイン関連の組織に属し、ドメインの専門知識を身につける、あるいは、3) ドメインの専門知識を有する人が、密に連携を取れるデータサイエンティストとタッグを組む、のいずれかが必要となる。以下の小節で、これら3つの場合についての詳細を述べる。

2.2 データサイエンスの学び直し

1) は、本稿の冒頭でも述べた、データを活用したい組織の人がデータ分析の「スキル」を身につけるため、学び直しを行う場合に相当する。しかしここでも、データ分析の統計や数学、分析手法に関する「知識」を学ぶことが、学び直しであると誤解するケースがある。実際に身に付けたいものが「知識」に留まるのではなく「スキル」である場合、車の運転やスポーツ、プログラミング等のスキルと同様に、どれだけの事例を経験したか、ということがスキルにつながる。すなわち、図1のデータ分析のプロセスを通じた一連の経験を、数を多くこなすことがスキル獲得に向けては重要であり、そのための分析手法は、最初は1つ（例えばテキストデータ内の高頻度語に着目する）だけでよく、データ分析のプロセスに慣れてきた段階で、分析手法のバリエーションを学んでいけば良い。これは、多くの時間が取れない、また理系出身でなく数学の基礎知識がない社会人の学び直しには、特に重要と考えられる。

2.3 データサイエンティストの就職

2) は、1) とは逆のアプローチで、大学などの機関でまずデータサイエンスを学び、その後企業に就職する場合に相当する。ドメインの知識なしにデータサイエンスのみを学ぶ場合の懸念としては、目の前のデータに対する分析結果の解釈が、どうしても浅くなりがちになる点が挙げられる。これは、単純にドメインに対する背景知識がないことに加え、データ分析に対する意欲が、ドメイン内の現場の人に比べると弱いことに起因する。そのためデータ分析に扱う題材を、分析者の興味や趣味に強く関わるデータ、とするなどにより、背景知識と意欲を補う工夫が必要と考えられる。

2.4 組織とデータサイエンティストとの連携

3) は、データ分析を専門とする企業や大学内のデータサイエンティストと、データ分析を必要とする組織が連携する場合に相当する。ただし、組織側がデータを分析側に渡し、組織側がその分析結果を受け取るだけの場合、データからわかる基本的かつ表面的な情報が得られるのみで、組織運営戦略への劇的なアドバイスになることは考えにくい。すなわち、組織側の人間が積極的にデータ分析に関わっていく姿勢を見せることが必要であり、図1のデータ分析のプロセスを組織側の人間が理解し、少なくとも「1. 分析目的」を明示的に伝え、「9. 結果の解釈」に関与することが必要と考えられる。単純には、組織側と分析側のデータ分析に関係するコミュニケーションの量が、データ分析の質に関わってくると言える。

いわゆるデータ分析の外注は、作業の独立性を高めるなど契約上のやり易さはあると考えられるが、データ分析の観点は組織側が持っている情報であり、データ分析プロセスにおける「気づき」の中で補完される情報となるため、分析プロセスの各項目について、適宜両者がコミュニケーションを取れる形で連携が実現されることが望ましい。そのためにも、2.2 節で述べた学び直しは、本節の組織側の人間にとっても重要と考えられる。

3. 情報アクセスと知識創発：知識に繋がられるデータとは？

本章では、データ分析に用いるデータを準備する方法について述べる。

3.1 蓄積されるデータの分析

分析の目的を考える前に、すでに手元に蓄積されているデータ（図1の「0. 蓄積データ」）が存在する場合がある。それらは、日常的に行われるメール等の電

子的なコミュニケーションの記録であったり、日々の業務の記録などのログデータとなることが多い。

手元に何らかのデータが蓄積されるにつれ、「このデータを分析すれば、何か新しい事実がわかるかもしれない」という考えが生じるのは自然なことで考えられる。しかし、この「何か」に通じる方向性を定めることなしに、具体的な分析を行うことはできない。例えば、「太平洋の海底には何かお宝が眠っているに違いないから、何かいいものを探ってもってきてほしい」と言われても困ることに似ている。「海賊の宝」を見つけない、「海底に眠る資源」を採掘しない、「珍しい深海魚」を発見しない、などの目的を定めることによって、現在のデータが目的に合致しているか、目的に対してデータ量や内容は十分か、目的を達成するための道具には何を使うか、それは用意できるのか、などを検討できるようになる。

そのため、多量のデータが日々蓄積されているというデータ分析の最初の段階においては、分析の目的をさまざまに考えることができるが、実際に分析の作業を行う前には、明示的に1つの目的を定める必要がある。

3.2 分析の目的の決定

我々の社会活動、社会生活は意思決定の繰り返しであり、データ分析の目的は、この「意思決定の根拠を得る」ことに他ならない。そのため、分析の目的の決定（図1の「1. 分析目的の決定」）に際しては、どのような意思決定を行いたいのかをまず考える必要がある。すなわち、「考えられる行動の選択肢の中から、望ましい結果につながる行動を選択すること」が、意思決定であり、その望ましい結果に繋がる根拠をデータ分析により探ることを目指す。

例えば、顧客の購買履歴データの分析により販売戦略を考える場合、「売り上げの増加」を分析の目的とする以外に、「顧客数の増加」や「顧客満足度の増加」などの目的を設定することも考えられる。しかし、分析を行うためには、分析の目的に対応する意思決定の根拠が含まれるデータを用意する必要があるため、仮に手元の蓄積データのみを活用して分析を行う場合には、そのデータが根拠を含むような分析の目的を設定する必要がある。

3.3 分析の目的に応じたデータの収集

分析の目的を達成する、すなわち意思決定の根拠となる内容を含むデータが、手元に蓄積されていない場合、あるいはより効果的に目的を達成する分析を行うためには、新たにデータを収集する必要がある（図1

の「2. データ収集」）。

どれだけの種類のデータを集める必要があるかについては、分析の目的により異なる部分はあるが、世の中の問題をコンピュータに解かせる際の、いわゆるフレーム問題（すべての情報をコンピュータに与えることができない）と同様に、およそ目的に強く関わると考えられるデータを順に集めていくことになる。

分析に用いるデータが、分析の目的に関わる知識を広くカバーしていれば、想定外の不測の事態を避けた分析を行えるようになる反面、分析が煩雑になるトレードオフが存在する。そのため、分析の目的が、主要な知識（一般的な事実や全体的な傾向）を把握するためのものであれば、目的に関わる主要なデータのみを対象とし、新しい知識（新事実や独自性のある傾向）を得るための分析においては、幅広くデータを集めるのが良いと考えられる。

3.4 データの価値の見積もり

前節で用意したデータの実際の分析に取り掛かる前に、目的を達成するための分析結果が得られるかを吟味して、データの価値を見積もることも大事である。すなわち、やみくもに分析を始めるのではなく、得られるはずと考えられる分析結果を1つでも2つでも想定して、結果に対する「仮説」を用意する。分析の目的に強く関わる「仮説」が多く立てられるデータほど、価値が高いデータとなる。

例えば、顧客の購買履歴データであれば、同時に購入される可能性が高い商品間の関係を抽出でき、「ある商品と同時にすすめると効果的な商品がわかる」という仮説が立てられる。このとき、具体的な商品名まで挙げる必要はない。またこのデータに、顧客の年齢層のデータがあれば、「特定の年代の人によく売れている商品がわかる」という仮説を立てることもできる。

仮説はもともと棄却される可能性があるものなので、「同時に売ると効果的な商品がなかった」「年代ごとによく売れている商品がなかった」という結果になることもある。しかし、より多くの仮説を立てられるデータであれば、いずれかの仮説に基づく有効な結果を得られる可能性が高まる。

この仮説を立てるためには、ある程度用いられる分析手法についての知識も必要となる。必ずしも分析手法の詳細なアルゴリズムを理解している必要はないが、主要な分析手法について、何を入力すると何が出力として得られるのか、言葉として理解できる分析手法の意味を知らなければ、適切な仮説を立てること、すなわちデータの価値を見積もることは難しい。

仮にデータの価値を見積もることができなくても、データの分析を行うことはできるが、分析の目的を意識した効率の良い分析ができなかったり、前節で述べた分析の目的に応じたデータ収集をうまく行えない可能性が生じると考えられる。

3.5 データ分析の前処理

分析に用いられるデータを集めた後は、分析ツールに入力できる形式にデータを変換するデータ整形（図1の「3. データ整形」）を行う。データの前処理は大きく、データ構造を対象としたものと、データ内容を対象としたものとに分けられる [2]。データ構造を対象とした前処理は、全データの中から分析に用いるデータを抽出することや、データ分析に用いるためにデータの平均や分散を計算することなどで、Fayyad らのデータマイニングプロセスの「データ選択」と「前処理」に相当する。またデータ内容を対象とした処理は、数値、文字、カテゴリデータなどの分析ツールによって指定されるデータの形式に変換を行って、データ形式を統一化することであり、Fayyad らのプロセスの「変換」に相当する。

留意したデータの形式とツールの入力形式は多種多様であるため、手作業によるデータ整形が必要であったり、データ整形のためのプログラムを記述する必要があるなど、地味に大きな労力を必要とするプロセスとして位置付けられており、データ分析の初心者がつまづきやすいポイントのひとつになっている。

最も手間がかからない入力形式としては、数値データの場合、表形式の2次元データを、カンマと改行で区切った CSV 形式のデータが用いられることが多い。しかしその場合においても、欠損値の補間、外れ値の除去、など値の調整を行ったり、値が実数値か離散値かの区別を与えたり、カテゴリデータとして文字列で表されているものを数値に変換する、などの処理が発生することがある。

またテキストデータの入力の場合は、1つのテキストデータをそのまま入力できると最も手間がかからない。しかしその場合においても、一文の終わりの句点の挿入や統一、文字コードの変換、分析に用いないストップワードの設定、文字でない記号の除去、などの処理が発生することがある。

煩雑な前処理を避けるための1つの方法としては、利用する分析ツールを固定し、そのツールに必要な入力を準備しやすい環境を整えることが考えられる。特に分析の初心者にとっては、まず1つのツールを使いこなせるようになり、要領が得られた時点で他のツールを利用するのが良いだろう。入力データの整形と並

行して行う作業に、データ分析に使用するツールを選択する（図1の「4. ツール選択」）必要がある。ツール選択においては、分析の目的を達成できるツールを利用する必要があることに加え、ツールの候補が複数考えられる場合は、ツールの入手とデータ整形に必要な費用および時間的なコストを考慮する必要がある。

複数のツールを使うことは大きな手間がかかるが、複数の観点から多様な分析結果を得るためには、さまざまなツールがひとつの環境内で用いられることが望ましい。例えば、著者らはテキストマイニングのための統合環境 TETDM[3] を構築し、フリーソフトとして公開している²。

4. マイニングと知識創発：解釈を集めるためのマイニングの出力とは？

本章では、データマイニングの出力を人間が認識しやすい形式で出力し、解釈を与えやすくする手法について述べる。

4.1 データマイニングの結果の獲得

データマイニングツールを用いて得られるデータ分析の結果とは、データ全体、または「条件」によって絞り込まれたデータについての事実や傾向を表す。

入力データの量が多い場合を初め、データからわかる多様な事実を得るためには、「条件」によってデータを絞り込む（図1の「5. データ絞り込み」）が必要となる。これにより、「ある条件を満たすデータにあてはまる事実」を結果として得られる。また、データマイニングツールを変えることで、出力となる事実の形式を変えることができる。

4.2 データマイニングの出力

データマイニング（図1の「6. データ処理」）は、相関ルールマイニング、クラスタリング、分類モデル構築に大別される。相関ルールは、「ある商品 A を購入した人は別の商品 B も購入する」というようなアイテム（データオブジェクト）間の関係をルール形式（ $A \Rightarrow B$ ）で抽出したものである。データセット内でよく共起するアイテム集合（頻出アイテム集合）を抽出した後、それを条件部（A）、結論部（B）に分割してルールを生成し、確信度やリフト値といったルールとしての良さを評価する指標に基づき有用なルールを抽出する。従って、マイニング結果の出力としては頻出アイテム集合のセットや、相関ルールのセットとなる。

2 TETDM サイト：(URL)<http://tetdm.jp>

クラスタリングは、データオブジェクトを与えられた類似度・距離に基づいて類似するいくつかのグループに分割する手法である。出力はデータオブジェクトの集合（クラスタ）のセットであり、加えて階層的クラスタリングの場合はクラスタ間の階層構造（包含関係）も出力される。クラスタ間の階層構造を可視化したものは樹形図（dendrogram）と呼ばれ、ルートをデータオブジェクトの全体集合、リーフをデータオブジェクトとする木構造として描画される。

分類モデル構築の目的は、データオブジェクトを既知のクラスに分類することであり、ニューラルネットワークや SVM（support vector machine）、決定木（decision tree）等が代表的である。所属クラスがわかっているデータオブジェクトを訓練データとして与えて分類モデルを構築し、所属クラスが未知のデータオブジェクトのクラスを予測する。予測のみが目的の場合、出力は入力されたデータオブジェクトに対する予測クラスとなるが、アルゴリズムの評価などを目的とする場合にはデータオブジェクトの正解クラスと予測クラスが出力され、それらの一致具合（誤分類）や予測クラスの境界などを分析する。

上記いずれの場合でも、個々のデータオブジェクトに対する詳細を確認することが一般的であり、その点ではマイニングの入力データも出力データの一部を構成するといえる。

4.3 代表的な可視化手法

可視化手法（図1の「7. データ可視化」）には多種多様なものが存在し、可視化対象となるオブジェクトの特性、および視覚的に確認したい内容に基づいて適切な可視化手法を選択すべきである。表1に、代表的な可視化手法を可視化対象に基づき分類したものを示す[4]。データオブジェクトを対象とした場合は、各オブジェクトが持つ属性を視覚的に確認することが目的となり、チャノフの顔グラフなどの図像的可視化、各属性をそれぞれ平行な軸としてオブジェクトを折れ線で描画する平行座標、軸を中心から正多角形上に配置するレーダーチャートなどが用いられる。

離散的データの場合、オブジェクト間の接続関係や上下関係などを確認することが目的となるため、ネットワーク構造や木構造の可視化手法が用いられる。連続的データの場合は、オブジェクト間の距離関係も重要となるため、多次元データを低次元空間に写像して可視化する力学的モデルや多次元尺度構成法、データの分布を可視化するヒストグラムや散布図が用いられる。また、時系列データも実数値の時間的変動が重要であるため連続的データの一種とみなせ、線グラフが

表1 可視化手法の分類

対象	可視化手法の例
データオブジェクト	図像的可視化, 平行座標, レーダーグラフ
離散的データ	ノードリンクダイアグラム, 接続行列, ツリーマップ
連続的データ	力学的モデル, 多次元尺度構成法, ヒストグラム・散布図, 線グラフ
その他	タグクラウド, ディストーション, 協調的マルチビュー

一般に用いられる。

その他の可視化手法としては、テキストデータを対象としたタグクラウド、Perspective Wall[5]などのように、詳細な情報（detail）と全体像（overview）を同一画面内に表示するために描画領域内でレイアウトや縮尺などを変化させるディストーション、複数のビューを連動させる協調的マルチビューなどが存在する。

データマイニングとの対応に関しては、相関ルールや階層的クラスタリングなどの出力はネットワーク構造、木構造に適した可視化手法が適用可能である。分類モデル構築の場合はデータの分類や誤分類データの把握のために、連続的データを対象とした可視化手法でデータ分布を確認することが多い。また、いずれの場合でも個々のデータオブジェクトの詳細を確認することが一般的であり、データオブジェクトの可視化が利用可能である。

5. 情報可視化と知識創発：解釈を集めるための情報可視化とは？

本章では、意味のある結果への気づきを促し、幅広い解釈を集めるための情報可視化の手法について述べる。

5.1 視覚的分析（Visual Analytics）

探索的データ分析は、データから新たな知見や仮説を得る目的で行われる作業であり、データの分布傾向などと言った全体像の把握、詳細な情報へのアクセスなどから構成される。情報可視化技術を用いた探索的データ分析は視覚的分析（Visual Analytics）と呼ばれる。視覚的分析において、データから何らかの知見を読み取るために、様々な角度からデータを眺めるビュー（可視化結果）を生成する作業を繰り返す。これらは、分析者にとって満足するビューを見つけるための探索的プロセスと捉えることができる。

視覚的分析において、ビューを眺めながら分析者が行う作業は、大きく以下の3点に集約される。

- 1) 可視化された情報の中で、興味ある部分、特徴的な部分に気づく。
- 2) 気づいた点について、その背後にある理由などについて考察する。
- 3) 「気づき」を積み上げ、何らかの知見を得る。

1) は、他と大きく異なる値をとるデータや、時系列データで大きく変動する時点、2 属性間にみられる関係性などに気づくことを意味する。これらの「気づき」を解釈し、意味づけるプロセスが 2) であり、センスメイキング (sense making) と呼ばれる。また、気づきは断片的なものであることも多いが、それを積み上げることで、最終的にデータ全体に対するより大きな、包括的な理解に至り、問題解決やアクションにつながる知見が得られる。3) はこのプロセスに対応し、洞察と呼ばれる。

5.2 協調的可視化によるデータ解析

探索的データ分析では多様な観点からデータを眺めることが重要であり、そのためには複数人の分析者が協調して探索を行うことが有効である。分析者毎に対象ドメインに対する知識や関心などが異なることが想定されるため、それぞれ異なる視点からデータを分析することになる。そのプロセス・結果を共有することにより、一人の場合よりも探索空間が広がり、データの深い理解につながる事が期待できる。

協調的なデータ分析の例として有名なのが、NameVoyager における可視化を共有してのコミュニケーションである。NameVoyager[6] は、米国における新生児の名前のトレンドを可視化して提供するサイトである。1880 年代から現在までの、米国における新生児数を名前毎に集計し、その時系列変化を積み上げ線グラフで可視化する。ダイナミッククエリによるフィルタリング機能を備えており、アルファベットの文字列をタイプすると、その文字列から始まる名前のみがグラフとしてリアルタイムに描画される。

興味深い現象として、NameVoyager のビューを引用しながら、利用者の間で様々な議論が交わされたことが報告されている。例えば、自分や知り合いの名前の人気の変遷を調べ、ユーモラスなコメントとともに紹介したり、珍しい名前を発見して発表したり、有名人の名前について気づいたことを発表したりといった事が行われた。ここで注目したいのは、NameVoyager のサイト自体はコメント投稿機能などは備えていない点である。すなわち、利用者たちは別の Web サイト、つまりブログやディスカッションフォーラムなどを利用してコミュニケーションを行っていたのである。その際、NameVoyager が生成するビューは、ダイナミック

図2 地域防犯活動支援システムの議論インターフェース

クエリとして入力する文字列を記載することで共有されていた。これにより、ある利用者により発見された気づきは他者と共有され、それに対し他者がコメントしたり、疑問点については回答が得られたりしてデータに対する理解が深まっていく様子が観測されたと言われている。

より積極的に協調的データ分析を支援する機能を備えたサイトも開発されている。Many Eyes[7] は、汎用的なデータ可視化・分析を目的とした Web サイトであり、利用者はデータセットのアップロード、ビューの生成・共有、それに基づく議論が可能である。

著者らは、地域防犯活動に関する議論を支援する協調的可視化システムを開発している (図 2) [8]。このシステムでは、児童達が学校周辺を歩いて調べた結果に基づき、オンライン地図を用いて地域安全マップ [9] を作成、共有可能としている。マップ上では児童が発見した、危険な場所や安全な場所について、写真と児童によるコメントが共有されている。地域安全マップの特徴として、危険な場所、安全な場所を「入りやすさ」、「周囲からの見えやすさ」の 2 つの観点から判断することで、児童達に危険な場所、安全な場所の判断基準を身につけさせることを目的としている。このシステムでもこの考え方に基づき、場所を登録する際にこれらの観点に関するアイコンを付与するようになっている。

また、共有されたマップを参照しながら、教員や保護者などが地域防犯活動について議論することを可能としている。前述のアイコンを場所に関する属性として指定し検索する機能が用意されており、共通した性質を持つ場所を集めたり、現在の議論と関連する他の場所を発見したりすることが可能となっている。

6. インタラクションと知識創発：解釈を集めるためのインタラクションとは？

本章では、図 1 の「8. 結果の収集」に関連して、多

様な解釈を集めやすくするための、試行錯誤の手段としてのインタラクション（図1の「5. データ絞込み」に通じるインタラクション）について述べる。

6.1 インタラクティブクラスタリングと可視化

インタラクティブクラスタリングは、対話的なデータ分析や情報アクセスなどにクラスタリングを適用したものであり、肺がんの潜在的要因の分析 [10] や類似映像バケット発見タスク [11] など、多様な用途で有用性が期待されている。

インタラクティブクラスタリングには、特徴選択 [10] や重み付け [11] など対話的な要素に関して多様な手法が存在するが、代表的なものの1つに制約付きクラスタリングが挙げられる。制約付きクラスタリングは対象データの分類に関して、人間の持つ常識や背景知識などに基づき制約を与える半教師あり学習であり、クラスタリングアルゴリズムは与えられた制約を満たすようにデータを分割する。代表的な制約表現に制約があり [12, 13], *must-link*, *cannot-link* の2種類の制約が一般に用いられる。*must-link* を付与されたオブジェクト対は同じクラスタに、*cannot-link* を付与されたオブジェクト対は異なるクラスタに分類されるべきであることを意味する。制約を考慮するように既存クラスタリングアルゴリズムを改良した手法が提案されており、K-means を拡張した COP-K-means [14] などが代表的である。COP-K-means では、通常の K-means と同様にクラスタ割当てを行う際、与えられた制約を満たす範囲で最も中心に近いクラスタにオブジェクトを割り当てる。制約を考慮するように階層的クラスタリングを拡張した手法も提案されている [15]。

クラスタリング手法に依存しないアプローチとして、*must-link* が与えられたオブジェクト間の距離は近く、*cannot-link* が与えられたオブジェクト間の距離は遠くなるように距離空間を更新する手法も提案されている。更新後の距離空間で求められるオブジェクト間距離に基づき、グラフカット [16] や K-means [17] など様々なクラスタリング手法を適用してクラスタリング結果を求める。

インタラクションの設計に関して、制約を分析者が逐次指定することを考えた場合、指定しなくてはならない制約数が多い場合はユーザの負荷が増大する他、付与された制約間に矛盾が生じる可能性が指摘されている [18]。これらの問題に対し、半自動的に制約を一括生成する手法も提案されている。著者らは、クラスタリング結果にユーザが施すグルーピング操作に基づき、制約を一括生成する手法を提案している

図3 ツリーマップを用いたインタラクティブクラスタリングインタフェースにおける分解操作の例：(a) 分割前、(b) 分割後

[3, 19]. 提示されたクラスタリング結果に対し、ユーザはクラスタの統合や分割といったグルーピング操作を行い、データ分割に関する自身の意図を反映させる。グルーピング操作履歴や階層的クラスタ構造を利用して制約を付与すべきデータ対を求め、制約付きクラスタリングに利用する。

グルーピング操作はクラスタの階層構造に対する操作とみなすことができるため、木構造の可視化手法であるツリーマップを採用したインタフェースを TETDM を用いて開発している。図3に開発したインタフェースのスクリーンショットを示す。矩形がクラスタに対応し、サイズは文書数に比例する。本インタフェースでは分析対象となる文書にラベルが付与されていることを想定しており、各クラスタ内で多数派となる文書のラベルに応じた色で矩形が描画されている。

クラスタは、対応する領域をマウスクリックすることで選択でき、直近に選択された2クラスタの情報が保持される。パネル下に配置されている統合ボタンを押すことで、選択中の2クラスタが統合される。分割ボタンを押すと、直近に選択したクラスタが2つに分割される。図の例では、左上の矩形に対応したクラスタに分割操作が適用されている。

6.2 セマンティックインタラクション

Endert らは、視覚的分析におけるインタラクションの設計においては、データ分析の専門家ではなくドメインの専門家を支援することを目的とすべきとしてい

る。具体的には、以下の原則を満たすセマンティックインタラク션을提案している [20, 21].

1. 視覚的メタファは統計モデルあるいは類似度尺度により生成され、分析者の空間的認知を支援する。
2. 空間的分析プロセスにおいて一般的なインタラクシオンに基づいて、視覚的メタファと意味的にインタラクシオンする。
3. ユーザとの意味的なインタラクシオンを解釈し、モデルのパラメータに写像する。
4. ビューの背後にある数学的モデルやパラメータの複雑さからユーザを保護する。
5. 分析プロセス全体を通じて発生するインタラクシオンを考慮することで、逐次的にモデルを学習する。
6. 更新されたモデル・パラメータを、視覚的メタファを用いて視覚的にユーザへフィードバックする。
7. 学習したモデルのパラメータはその後の分析において再利用する。

視覚的分析においては、何らかのモデルに基づくデータ分析結果がビューとしてユーザに提示される。例えば、データの空間的配置は、距離学習により得られた統計モデルにより決定される。提示された結果に対し分析者が何らかの修正を行いたい場合、従来型のインタラクシオンではコントロールパネルなどを操作し、モデルのパラメータを直接フィードバックする必要がある。これに対しセマンティックインタラクシオンでは、空間的メタファを用いて提示されているデータに対し分析者が操作を行い、その操作に込められた意図を解釈してモデルへのフィードバックとして利用する。この時、分析者が行う操作は特殊なものではなく、自然で慣れたものを採用する。例えば、分析者が文書を画面上のある位置に移動させて固定した場合、場所に意味を与える行為と解釈する。固定された文書に他の文書を重ねた場合、両文書に共通するエンティティ（キーワード）の重要度を高く更新する。また、文書に注釈を付与した場合、注釈からエンティティを抽出して重要度を付与するとともに、エンティティを共有する文書間にバネを生成し、近い位置に配置されるようにする。

データマイニングなどのプロセスを完全な自動化として捉えるのではなく、分析者である人間を含めるコンセプトを *human in the loop* と呼ぶ。特に、分析者が問題領域の専門家である場合、機械学習モデルなどについて十分な知識を持っていないことが想定されるため、セマンティックインタラクシオンのようにイン

タラクシオン設計における配慮が重要である。また、モデル自体もドメイン専門家が理解しやすく、知識が与えやすいようなパラメータを持つように設計すべきとの指摘もある [22].

7. 知識創発：集めた解釈をどのように知識につなげるか？

本章では、前章までの環境により集められる分析結果について、それを解釈に変換する作業（図1の「9. 結果の解釈」）、ならびに集められた解釈の集合を、整理、統合する中で、新たな知識に結びつけるプロセス（図1の「10. 解釈の整理」）について述べる。知識整理に基づく発想法としては、KJ法 [23] が古くから知られており、まとめた情報を1つずつ記録し、それらを似たもの同士のグループにまとめることが効果的となる。本章では、このまとめるという操作を主とする知識創発のプロセスについて詳しく見ていく。

7.1 分析結果の意味づけによる解釈

分析結果の解釈とは、分析結果として得られた事実や傾向に、そのデータのドメインについての背景知識や世の中の情勢など、データ化する際に捨象された情報を加えながら、結果が表す意味を読み解くことである。

データサイエンティストは、さまざまなツールを使い分けるなど、データを分析する作業自体はこなすことができるが、集めた結果の解釈においては、データが対象とするドメインの知識がないと、十分な解釈を行えない。

データが対象とするドメインの専門的知識を有する人、一般的にはデータに関わる現場の人であれば、得られた結果をもとに、十分な解釈を行うことが期待できる。しかし、結果の収集までをデータサイエンティストが行い、解釈を現場の人が行えば良いかというと、必ずしもそうはならない。結果を収集する過程においては、情報を絞り込む、プロセスがあるが、その時に現場の人でなければ思いつかない、絞り込みの観点が生じる可能性がある。そのため、ドメインの知識を有する人が、結果の収集と結果の解釈の両方に関わってることが望ましい。

7.2 解釈集合を整理するプロセス

多くの解釈があった時に、それらを別々のものとして扱うのではなく、複数の解釈をまとめたシンプルな解釈を逐次生成することで、汎用的で応用可能性の高い知識を得ることができる。

この複数の解釈をまとめるプロセスは、1) 複数の

図4 解釈の統合と論理の飛躍

図5 解釈の再帰的統合

解釈の共通点を見つける、2) 見つけた共通点に基づいて、複数の解釈を一般化した新たな統合解釈を生成する、という2つのステップからなる。例えば、スーパーの販売履歴のデータを分析した結果の解釈として、「秋には果物がよく売れる」「秋にはゼリーがよく売れる」という2つの解釈が得られたとする。この2つの解釈の共通点として、たとえば「食後のデザート」を挙げた上で、「秋には食後のデザートがよく売れる」という2つの解釈を統合した解釈を得ることができる。

この解釈をまとめるプロセスでは、少なからず論理の飛躍が生じる。解釈の統合と論理の飛躍との関係を図4に示す。図の左側は、まとめた解釈とほぼ必要十分な関係にある統合解釈Aであり、こちらは論理の飛躍が少ない。また図の右側は、まとめた解釈を拡大して一般化した統合解釈Bであり、こちらは論理の飛躍が大きくなる。すなわち論理の飛躍の大きさは、まとめる元の解釈集合の大きさと、統合解釈の大きさととの差（図4の解釈1から3の楕円の面積の集合和と、統合解釈の楕円との面積の差）として、式(1)で定義される。ただし、解釈の大きさとは、解釈が当てはまる対象の広さ、を表すとする。

（論理の飛躍の大きさ）

$$= (\text{統合解釈}) - (\text{解釈1}) \cup (\text{解釈2}) \cup (\text{解釈3}) \quad (1)$$

たとえば先の例の場合、「果物」と「ゼリー」を「食後のデザート」としてまとめることには大きなずれはないと考えられるが、これら2つだけの解釈を「甘いもの」や「食べ物」などのより一般化したものでまとめると、拡大解釈になる恐れが大きくなる。

堅実な（論理の飛躍が少ない）解釈を繰り返す場合には、信頼度が高い結果が得られるため、データが得られた背景を表す現状理解に役立てられる。一方で、論理の飛躍が大きい解釈を繰り返すと、信頼度は低くなるが、意外な、突飛な、可能性を秘めた知識を得られる可能性が高まる。

解釈集合の整理は、この解釈の統合を再帰的に繰り返すことで行われる（図5）。得られた全ての解釈を、1つの解釈にまとめられる（図5の統合解釈Xを得る）まで繰り返してもよいし、主要な複数の解釈にまとめられた（図5の統合解釈A,B,Cを得た）時点で終了としても良い。いずれにしても、分析者が活用可能な知識が得られたと感じられることが重要となるため、解釈の統合を繰り返しても活用可能な知識が得られない場合は、統合時にまとめる解釈の組み合わせを変えてみたり、統合解釈の論理の飛躍の程度を変えてみたり、時には新しい解釈を追加するためにデータ分析のプロセスに立ち戻ることも必要と考えられる。

返すことで行われる（図5）。得られた全ての解釈を、1つの解釈にまとめられる（図5の統合解釈Xを得る）まで繰り返してもよいし、主要な複数の解釈にまとめられた（図5の統合解釈A,B,Cを得た）時点で終了としても良い。いずれにしても、分析者が活用可能な知識が得られたと感じられることが重要となるため、解釈の統合を繰り返しても活用可能な知識が得られない場合は、統合時にまとめる解釈の組み合わせを変えてみたり、統合解釈の論理の飛躍の程度を変えてみたり、時には新しい解釈を追加するためにデータ分析のプロセスに立ち戻ることも必要と考えられる。

7.3 解釈集合を整理するプロセスの支援

解釈集合を整理するプロセスは、基本的に人間が行うことになるが、何の手がかりもなしに複数の解釈をまとめることは、時に手間がかかったり、分析者がもつ背景知識に依存する結果になることがある。そこで本節では、この解釈集合を整理するプロセスをコンピュータが支援する方法について述べる。

まず、まとめる解釈を選択する必要がある。すなわち、共通点が見つけられる解釈を選ぶ必要があり、必然的に類似する解釈同士を選択していくことになる。そのため、現在得られている解釈間の類似度を測り、類似する解釈同士をリンクで繋いだり、近くに配置することで支援ができる。

次に、統合解釈を与えるためには、解釈間の共通点を探る必要がある。そこでは、各解釈のキーワードについて、シソーラス上の上位概念にあたる言葉を提示したり、検索エンジンを用いて、両キーワードを含むWebページ内に現れやすい単語を提示する支援が考えられる。これは、分析者がもつ背景知識に依存しない解釈を促すためにも役立てられる。

また、特に新たな発想を促す目的では、まとめる解釈を選択する際に、類似度が高い解釈同士を選択するのではなく、類似度があまり高くない解釈のペアを提示し、それらの統合解釈を促す方法も考えられる。類

表2 獲得したい知識と分析手法

獲得したい知識	利用データ	利用ツール	解釈の方針
主要な知識	主要なデータ	メジャーなツール	堅実な解釈
新しい知識	幅広いデータ	マイナーなツールを含む	論理の飛躍を含む解釈

似度が高くない解釈をまとめる際には、論理の飛躍が大きくなる可能性が高くなるため、堅実な解釈を望む分析者にあえて提示することで、発想が促される可能性が生じる。

これら、解釈間の類似度の測定と、共通点抽出の自動化ができるようになれば、コンピュータによる統合解釈の自動生成も、1つの選択肢として与えられる日が来ると考えている。

7.4 創発される知識の価値の推定

本節では、最終的に得られた統合解釈から得られる知識の価値の推定について述べる。

価値判断の基準としては、研究費の公募と同様に、アイデアの価値を評価することに類すると考えられる。そのため、得られた知識の「適用可能性」「適用時の効果」「革新性と波及効果」が、主な価値判断の基準になる。

「適用可能性」は、どんなに素晴らしい知識であったとしても、それが実際に使えないと意味がないため、実際に適用可能かどうかを測る。「適用時の効果」は、知識を適用した時の効果の大きさを測る。「革新性と波及効果」は、目の前の適用対象への効果とは異なる視点となるが、その知識が今までにはない新しさを有する時に、その考え方が別のところにも応用され、広く展開される可能性を測る。

「適用可能性」と「適用時の効果」については、解釈集合を整理するプロセスにおいて、基本的に論理の飛躍が少ない方が、実現の可能性が高く、確実な効果が期待される知識につながると予想される。逆に「革新性と波及効果」については、当たり前の知識では革新性が生まれなため、論理の飛躍を大きくすることが必要と考えられる。すなわち、解釈を統合する際の論理の飛躍の程度を測ることができれば、そこから創発される知識の価値を見積もる1つの指標として活用できる。これは、統合前の解釈と、統合後の解釈との間に、どれくらいのギャップがあるかということに相当し、前節で述べた統合解釈の自動生成を行う際の指標が用いられると考えている。

7.5 知識創発の目的に応じたデータ分析の方針

本節では、知識創発の目的に応じたデータ分析の方

針について述べる。

知識創発によって最終的に獲得したい知識は、大きく分けて、データの対象となっているドメインの現状把握や、知識を他人に伝えることを目的とする「主要な知識」の獲得と、データを手掛かりとした新しい分野の開拓、新アイデアの発見を目的とする「新しい知識」の獲得とに分けられる。

この大別される2つの目的に応じた、利用データ、利用ツール、解釈の方針を表2に示す。利用データについて、主要な知識を獲得するためには主要なデータを集めれば良いのに対して、新しい知識を獲得するためには、少しでも手がかりになりそうなものを網羅的に幅広く集めることが望まれる。

利用ツールについて、主要な知識を獲得するためには、統計的、数学的な確立された、あるいは広く用いられている、いわゆる信頼度が高いツールを用いれば良いのに対して、新しい知識を獲得するためには、条件付きで有効に働くような、必ずしも信頼度が高いツールを含めて用いることで、幅広く手がかりを集めることが望まれる。例えばテキスト分析であれば、前者は高頻度の単語にのみ着目すれば良いが、後者のためには、低頻度の単語にも着目する必要がある、多くの低頻度の単語の中から有効な単語を抽出する方法は、ツールの特性による影響を受けやすい。

最後に解釈の方針については、前節で述べた通り、主要な知識を獲得するためには堅実な解釈が望まれ、新しい知識を獲得するためには、論理の飛躍を含む解釈が望まれる。論理の飛躍を行う際には、その飛躍の方向性がさまざまに考えられるため、解釈の統合における試行錯誤が必要になる。

いずれにしても、新しい知識を獲得するためには、手がかりを幅広く集めた上で、試行錯誤に十分な時間と手間をかける意欲を持つことが必要となる。

8. 知識創発プロセス支援の具体例

本章では、図1後半の知識創発プロセスを支援する枠組みの例として、テキストマイニングのための統合環境 TETDM[3]で行っている支援について述べる。

TETDMの画面例を、図6に示す。ウインドウは大きな3つのエリアから構成されており、各エリアにはTETDMに用意されている処理ツール約50種類と可

図6 TETDMの画面例

視化ツール約 30 種類の中から、1 つのペアを指定して割り当てることができる。処理ツールには、文章要約や文書クラスターリングなどのツールに加え、単語の頻度計算、主語抽出、長文抽出などの基本的なテキスト処理ツールも含まれる。また可視化ツールには、処理ツールに特化した可視化ツールや、テキスト表示、表形式表示、ネットワーク表示などの基本的な可視化を行うツールが含まれる。

そのため、図 1 の「4. ツール選択」における多様な結果を集めるためのインタラクションが可能となっている。また、「5. データ絞込み」のために、各ツールで着目している情報（たとえば文章要約の観点語）を含むデータに絞り込む機能があり、逐次的なインタラクションを支援している。

8.1 TETDM における結果の収集と解釈の支援

TETDM には、ツールが出力する結果を集めるための結果と解釈の登録インターフェース（図 7）が用意されている。同ウインドウには、気になる「結果」を入力するテキストフォームと、その結果の意味を「解釈」として登録するテキストフォームが用意されている。例えば、浦島太郎のテキストを入力した時に、図 6 中央部のキーワードが出力されたとする。そこで気になるキーワードとして、「人間」「大事」に着目した場合、結果として「人間や大事という単語がキーワードとして出力された」と記入する。そして、これらの単語がキーワードとして出力された意味を考え、解釈として、例えば「人間にとって大事なものの話がある」と記入する。

このような結果と解釈を、さまざまなツールの出力を見る中で、積み重ねていくことができるようになっている。同時に、登録時の画面をキャプチャする機能や、登録時のパネル構成を記憶する機能があり、登録した結果の一覧を表示する画面においては、結果と

図7 TETDMの結果と解釈の登録インターフェース

解釈を登録した時の状況を再現することも可能になっている。

この結果の解釈と登録の作業は、支援インターフェースを有しない環境においては、分析者が頭の中だけで行ったり、紙やエディタに記録するという方法がとられることが多い。しかし、別媒体を用いるとその手間によって思考が途切れがちになったり、なぜその解釈を考えたのか、解釈の根拠となるデータを再度確認したり、解釈の根拠データを提示する必要がある場合に手間がかかる可能性が生じる。最初から重要な結果がどれかわかっていれば、それについて細かく記録することもできるが、知識創発の前半においては、可能性を秘めた結果を発散的に幅広く集めることが求められるため、手間をかけずに必要な情報を残せる、データと思考を結び付けられる環境を利用することが望まれる。

8.2 TETDM における解釈の整理の支援

TETDM には、前節で集めた解釈を整理統合するための知識創発インターフェース（図 8）が用意されている [24]。

インターフェース内には、前節の結果と解釈の登録インターフェースを用いて集められた解釈が並べられており、これらに対して 7.3 節で述べた再帰的な統合を繰り返すことで、統合解釈を得ることができる。

インターフェースにおいては、関連する複数の解釈をマウスでクリックして選択した上で、選択した解釈の「共通点」を記入するテキストフォームと、その共通点から考えられる「統合解釈」を記入するテキストフォームが用意されている。この統合解釈は、多段に渡って行うことができるように複数のフォームが用意されており、三段論法の要領で、最初に行った統合解釈から考えられる解釈を記入することもでき、これに

よって解釈における論理の飛躍の程度を調整することもできる。

例えば、図7において、「ファンがアイドルに求めること」を分析する際に、「非日常的で刺激的な機会を体験できる」という解釈と「日常のストレスを忘れることができる」という解釈を統合する際には、共通点として「日常からの脱却」を挙げ、統合解釈1として「非日常の体験機会となること」を記入することができる。この統合解釈1は、元の解釈との関連度が高く堅実な解釈に相当する。そのため、アイドルの運営会社が必ず抑えるべき項目を検討する際にはこの解釈を使うことが望ましい。しかし分析の目的が、新しいアイドルの売り出し方を考えたい場合においては、より拡大した解釈を与えることも必要となる。そこで統合解釈1を発展させて、「見たこともない能力を持つ」という設定を与える」という統合解釈2を記入し、さらに「魔法が使える」という統合解釈3を記入して、論理の飛躍を実現することもできる。

この多段解釈では、元の統合解釈を拡大することも可能であるが、むしろ方向性を定めた統合解釈の絞り込みを可能にしておき、解釈の統合時に、解釈の拡大と方向性を定めた絞り込みの両方を行えるようにすることで、任意の解釈にたどり着ける可能性をもたせている。すなわち、解釈の統合は解釈集合を収束させるのに対して、多段解釈は解釈を発散させる方向にも導くものとなるため、多段解釈を用いることは、集めた解釈を単に収束させるのではなく、多少の発散を許しながら収束させることに相当する。この過程は、推理小説における推理と同様に、問題の解決となる手がかりを集めながら、複数の手がかりが示す状況をまとめるとともに、そこから推論される事柄を導くことに喩えられる。

インタフェースにおいては、統合解釈を記入した上で、最下部の「統合」ボタンを押すと、選択された統合をまとめた新たな解釈が、選択した解釈に置き換わって表示される。その置換した解釈を用いて、再帰的に解釈の統合を繰り返すことができる。

図7のインタフェースは、解釈の統合を支援するために、解釈間で用いられている同じ単語の色を変えて表示し、解釈間の関連度に応じて散布図形式で表示する機能、統合の履歴を保存することで、統合過程の再現や統合のやり直しができる機能を有する。これによって、さまざまな統合の可能性が検討できる統合の試行錯誤を支援している。また、明示的な因果関係としての解釈を生成したい場合、統合解釈に「原因」または「結果」のラベルを与えることができ、1つの統合解釈を「原因」と「結果」に分割することもできる。

図8 TETDMの知識創発インタフェース

例えば先の例においては、統合解釈「魔法が使える」に「原因」のラベルを与え、同時に統合解釈「ファンが元気になる」を生成して「結果」のラベルを与えることで、「魔法が使える→ファンが元気になる」という因果関係の知識として扱うことができるようになる。その後、「原因」のラベルが与えられた解釈同士、「結果」のラベルが与えられた解釈同士を統合していくことで、より大局的な因果関係を与える解釈の生成に結び付けられる。

9. まとめ

本稿では、知識創発につながるデータ分析のプロセスについて述べるとともに、データ分析を支援する環境について述べた。データ分析は必ずしも専門的な知識を必要とはせず、我々が普段の日常生活で行っている意思決定の類題にすぎない。世の中の多くの人が、データ分析の本質を理解し、目の前のデータを効果的に扱えるようになることを期待する。本稿がその一助となれば幸いである。

参考文献

- [1] U. Fayyad, G. P. Shapiro, and P. Smyth: "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, Vol.17, No.3, pp. 37-54, 1996.
- [2] 本橋智光: 前処理大全—データ分析のためのSQL/R/Python 実践テクニク—, 技術評論社, 2018.
- [3] 砂山渡, 高間康史, 徳永秀和, 串間宗夫, 西村和則, 松下光範, 北村侑也: "統合環境 TETDM を用いた社会実践," 人工知能学会論文誌, Vol.32, No.1, pp. 1-12, 2017.
- [4] 高間康史: 情報可視化, 森北出版, 2017.
- [5] J. D. Mackinlay, G. G. Robertson, and S. K. Card: The perspective wall: detail and context smoothly integrated, pp. 173-176, 1991.
- [6] J. Heer: "Socializing Visualization," *CHI2006 Workshop on Social Visualization*, 2006.
- [7] F. B. Viegas, M. Wattenberg, F. Ham, J. Kriss, and M. McKeon: "ManyEyes: a Site for Visualization at Inter-

net Scale,” *IEEE Trans. on Visualization and Computer Graphics*, Vol.13, No.6, pp. 1121-1128, 2007.

- [8] 高間康史, 瀬尾優太: “可視化表現共有型掲示板システムによる地域防犯活動議論支援,” 知能と情報, Vol.21, No.3, pp. 316-326, 2009.
- [9] 小宮信夫: 地域安全マップ作製マニュアル (改訂版) - 子どもと地域を犯罪から守るために-, 東京法令出版, 2006.
- [10] D. Guo: “Coordinating computational and visual approaches for interactive feature selection and multivariate clustering,” *Information Visualization*, Vol.2, No.4, pp. 232-246, 2003.
- [11] 木下幸裕, 新田直子, 馬場口登: “適合性フィードバックを用いた映像パケットのインタラクティブクラスタリング,” 電子情報通信学会論文誌, Vol.J88-D-II, No.8, pp. 1702-1711, 2005.
- [12] S. Basu, A. Banerjee, and R. J. Mooney: Active semi-supervision for pairwise constrained clustering, *SDM-2004*, pp. 333-344, 2004.
- [13] D. Klein, S. D. Kamvar, and C. D. Manning: “From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering,” *ICML’02*, pp. 307-314, 2002.
- [14] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl: “Constrained K-means clustering with background knowledge,” *ICML’01*, pp. 577-584, 2001.
- [15] S. Miyamoto, and A. Terami: “Constrained agglomerative hierarchical clustering algorithms with penalties,” *FUZZ2011*, pp. 422-427, 2011.
- [16] 岡部正幸, 山田誠二: “制約付きグラフカットによる逐次クラスタリング,” 人工知能学会論文誌, Vol.27, No.3, pp. 193-203, 2012.
- [17] 岡部正幸, 山田誠二: 制約付き距離学習による文書クラスタリング, *JSAI2009*, 1B1-3, 2009.
- [18] 青島傳隼, 福田直樹, 横山昌平, 石川博: “マイクロブログを対象とした制約付きクラスタリングの実現,” *DEIM Forum 2010*, B1-3, 2010.
- [19] 北村侑也, 高間康史, 梶並知記: “グルーピング操作に基づく対制約一括生成手法,” 人工知能学会論文誌, Vol.31, No.1, pp. 1-9, 2016.
- [20] A. Endert, P. Fiaux, and C. North: “Semantic Interaction for Visual Text Analysis,” *CHI’12*, pp. 473-482, 2012.
- [21] A. Endert, P. Fiaux, and C. North: “Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering,” *IEEE Trans. Visualization and Computer Graphics*, Vol.18, No.12, pp. 2879-2888, 2012.
- [22] S. J. Taylor and B. Letham: Forecasting at Scale, *PeerJ Preprints*, 2017: doi:10.7287/peerj.preprints.3190v2
- [23] 川喜田二郎: 発想法 - 創造性開発のために -, 中央公論社, 1967.
- [24] 西山知志, 砂山渡: “分析結果とその解釈の統合を支援する知識創発インタフェース,” 第12回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会資料, pp. 41-46, 2016.

(2018年10月19日 受付)

[問い合わせ先]

〒522-8533 滋賀県彦根市八坂町 2500

滋賀県立大学工学部

砂山 渡

E-mail: sunayama.w@e.usp.ac.jp

—— 著 者 紹 介 ——

すなやま わたる
砂山 渡 [正会員]

1995年大阪大学基礎工学部制御工学科卒業。1997年同大学院博士前期課程修了。1999年同大学院博士後期課程中退。同年同大学院助手。2003年広島市立大学助教授。2007年同准教授。2016年滋賀県立大学教授。現在に至る。博士(工学)。人間の創造活動を支援する研究に興味を持つ。人工知能学会, 電子情報通信学会, 言語処理学会, IEEE, 各会員。

たかま やすふみ
高間 康史 [正会員]

1994年東京大学工学部電子工学科卒業。1999年同大学院工学系研究科博士課程修了。1999-2002年東京工業大学大学院総合理工学研究科助手。2002-2005年東京都立科学技術大学助教授。2005-2014年首都大学東京システムデザイン学部准教授。2014年より同教授。博士(工学)。Web Intelligence や情報可視化, 知的インタフェースの研究に従事。主要著書は「インテリジェントネットワークシステム入門」(コロナ社)。IEEE, 情報処理学会, 電子情報通信学会, 各会員。