ビッグデータの深層を斬りだす ～学習理論がえぐる「ディープ・ナレッジ」

1990年代の初頭にIBMの研究部門により提唱されたデータ・マイニングは、膨大なデータの中から価値ある知識を発見するための汎用技術として、マーケティング分析やセキュリティー、障害検出など、幅広い分野への応用が急速に進展しています。これまでのデータ・マイニングは、販売履歴の分析による「併売分析（同時購入する確率の高い商品の把握）」など、表面的な関係を提示してビジネスに役立てることが中心でした。しかし現在のビジネスにおいては、大量データからより深い知見を見いだし、次の意思決定につなげることが求められるようになっています。

今回の特別インタビューでは、データ・マイニング研究の第一人者である東京大学大学院 情報理工学系研究科 創造情報学専攻 教授 山西健司氏に、学習数理情報学の中核となるレイテント・ダイナミックスとディープ・ナレッジの世界について話をうかがいました。

学習数理情報学の中核となるレイテント・ダイナミックス

―― 現在の研究領域についてお聞かせください。

山西氏:私の研究領域である「学習数理情報学」とは、データの背後に潜む関係性や変化を捉えて価値ある情報を引きだすための研究です。領域の名称に含まれている「学習」とは「機械学習」を指しており、機械学習はデータの中から知識を獲得して将来に生かすことを目標としています。その知識は、大量データの中から特定の用途に適した数理モデルを抽出(学習)することで得られます。機械学習の理論的側面である学習数理情報学では、基礎理論である「情報論的学習理論」と実践的応用である「データ・マイニング」を機械学習の研究推進の両輪と位置付けています。

情報論的学習理論は、「学習とは、与えられたデータから、データを最も短く記述できるモデルを見つけることである」という、データ圧縮の原理に基づいた理論です。この原理を「MDL(Minimum Description Length)原理」と呼びます。これは情報理論の一大原理です。MDL 原理によれば、「情報を最も簡潔に記述する(圧縮する)メカニズムの中にデータを説明する本質が隠れている」と考えることができます(**図 1**)。このような視点に立つと、ほとんど全ての学習問題を統一的に理解することができるようになります。MDL 原理は、何かを説明するためにはより単純で、不要なものを削ぎ落とした論理を用いるべきであるという「オッカムの剃刀」という考え方を厳密化したものです。一方、データ・マイニングは、機械学習の応用的側面に注目した学問分野の一つです。大量のデータの中から構造的な知識を獲得し、価値に変えることを研究します。

私は、情報論的機械学習を実世界に応用する際に特に注目すべき概念として、「レイテント・ダイナミックス(Latent Dynamics):潜在的な構造変化」を提唱しています。レイテント・ダイナミックスとは、併売分析が対象とするような表層的な知見ではなく、データの背後にある“潜在的”な構造の変化や動きのことです。これにこそ貴重な情報が含まれているという観点に立って、大量データから真に価値ある情報を見つけ出すことを目指しています。この場合も MDL 原理に基づいてレイテント・ダイナミックスを発見できるのです(**図 2**)。

学習数理情報学はすでに多くの成果を上げており、さまざまな分野に貢献しています。具体的な応用分野として、私が前職の研究所に所属していた頃には「テキスト・マイニング」と「異常検知」という、大きく 2 つの領域に注力して推進してきました。当時は研究の実用化を強く求められる時代でしたので、異常検知では、どこよりも先駆けてソリューションを実現したいと思い、パイオニアとなるべく研究に取り組んでいました。例えば、セキュリティー・ログに異常値を見つけた場合、そこから犯罪や攻撃や詐欺のリスクを探し出し、予防するという実用性や利用価値の高いソリューションの創造を目指していました。

直近では、マーケティング分野への活用、教育データの解

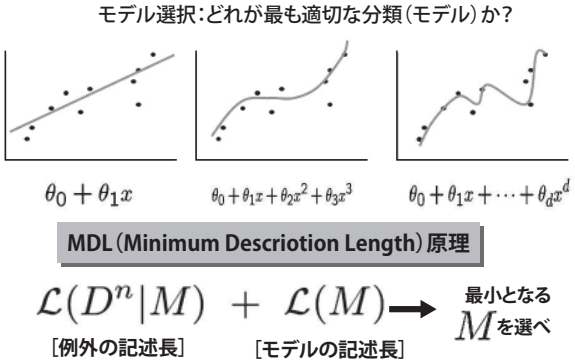

図 1. MDL 原理に基づく学習

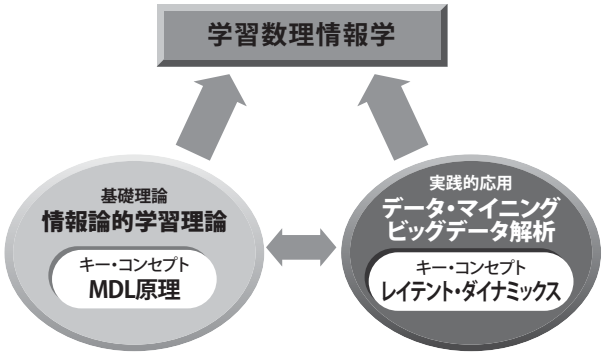

図 2. 学習数理情報学における両輪

P R O F I L E

[やまにし・けんじ]

1987 年、東京大学大学院工学系研究科計数工学専門課程修士課程修了。2008 年まで、主席研究員、データマイニング技術センター長として NEC 中央研究所にて機械学習、データ・マイニング、テキスト・マイニングの研究開発に従事する。途中、1992 年に東京大学大学院工学系研究科 博士号取得。2009 年 1 月より現職。

析、生命科学への応用など、より幅広い分野への次世代データ・マイニングの適用を推進しています。

純粋な学問的研究から実世界の要請に応える研究へ

―― 学習数理情報学に関心を持ったきっかけは何だったのでしょうか。

山西氏：学生のときから情報理論や符号理論の研究をしており、それを継続するつもりで企業の研究所に就職しました。しかし私が就職した1980年代は「符号理論は終わった。これからはAI（Artificial Intelligence：人工知能）の時代である」という風潮でした。当時、人工知能はロジックをたどる推論の研究が主流の研究分野でしたが、自分なりに人工知能の切り口を定めたいと思い、「機械学習」に着目しました。先にも述べたとおり、学習は情報理論的な考え方（特に、データ圧縮）が基礎原理をなしています。いったんそうした視点に立つと、情報理論や統計科学、計算機科学など、さまざまな理論が学習を通じて融合していく絵を描くことができました。そこで、これぞ自分の進むべき道だと信じて機械学習の理論研究を10年程度集中的に取り組みました。当時はデータ・マイニングという言葉も、インターネット技術もありませんでしたので、大量データをビジネスに生かすというよりも、機械はどこまで学習できるのかという学問的な関心が中心でした。

日本の1980年代は、基礎研究においても欧米に倣うばかりでなく日本独自の研究を進めようとしている時代でした。企業でも基礎研究を推進することができました。しかし景気が低迷した1990年代の半ばから、社会全体の雰囲気として経営資源を基礎研究よりも事業に直結する実用的な研究にシフトすべきという傾向が強くなりました。

そのような状況の中、それまで蓄積してきた機械学習の知見とノウハウを実用に転じる最適な分野として、私はデータ・マイニングに注目しました。データ・マイニングは、ちょうど1990年代の初めに出てきた言葉であり、注目度の高まっていた分野でした。私はこうした社会の要請に応える形で、実ビジネスに実装するための研究へと方向転換したのです。世界的に見た場合、データ・マイニングという大きなくくりの中では後発になってしまいましたが、その中でも異常検知とテキスト・マイニングに関しては先鞭をつけ、実績を挙げていきました。

重要なのは分析結果を次のアクションにつなげること

―― この10年間でデータ・マイニングは、どのような進化を遂げているのでしょうか。

山西氏：データ・マイニングには「おむつを買う人はビールも買うことが多い」という併売分析の有名な事例があります。最初期のデータ・マイニングはこのような知識を機械的に見つけることを目標としていました。しかし現在は、仮に「こうしたら売り上げが伸びる」というような相関ルールを見いだしてマーケティング担当者に提案しようとしても響きません。「そんなことは知っている。だからどうすればよいの?」と言われるのが実情です。併売分析のような前向きのデータ・マイニングは、IT予算が潤沢にあるとき、あるいは世の中がポジティブなムードで成長しているときには歓迎されますが、ネガティブなムードの場合には財布のひもはしまり、投資は後回しにされる傾向にあります。

そのような状況下においても、企業が欲しがるソリューションとはどのようなものかを考えたことが、異常検知の研究を始めたきっかけでした。重要なのは、得られた分析結果から次のアクションにつながる情報を提供することです。データ・マイニングによる異常検知によると、「もしその異常を発見できなかったり、放置したりすれば、どのくらいの犯罪やリスクにつながる、どの程度のコストにつながる」というように、それに対応しない場合のコストや脅威を具体的に示すことができます。こうしたリスク管理は、セキュリティーやコンプライアンスに関わるものとして企業の経営層の方々にも理解していただきやすい重要な領域です。

異常検知には、大量データから異質なデータを捉える「アウトライヤー（外れ値）検出」という分野があります。高次元データや時系列のデータの中から外れ値がどのように出るかを調べていくと、異常の出方にはあるパターンがあり、これまで蓄積した学習理論の力でそれを定式化できることがわかりました。こうした取り組みの成果を企業向けの異常検知ソリューションとして体系化していったのです。

体系化に伴い、例えばクレジットカード詐欺や遠回りによるタクシー料金の過剰請求など、異常検知のソリューションによってさまざまな事例が見つかり、学問的にも、機械学習の新たな研究領域として注目を集めるようになりました。

異常検知の考え方は教育データ・マイニングの分野にも応用できます。例えば、学校や塾などには膨大な量の学生の試験データが時系列に蓄積されています。このデータを分析することで、生徒の能力がどのように向上していくかを明らかにすることができます。成績データは、ある時点での試験の得点という意味しか持たないかもしれません。しかし時間という横串を使って時系列に分析することで、「この生徒はどの段階でスキルを身に付けたのか」「学びの成果として、どのような問題を解けるようになったのか」というスキルの成長に関する変化や異常を分析できます。この「スキル」というのはデータの表層からは直ちには得られない潜在的な概念です。ここでの「スキルの変化」はまさしくレイテント・ダイナミックスの典型であり、このような分析を得ることができたら、個々の生徒に最適化した教育システムやカリキュラムを提供することが可能になるでしょう。

「ディープ・ナレッジ」の発掘が次世代データ・マイニングの鍵

―― 企業の経営者や分析担当者がデータ・マイニングを活用するために取り組むべきことはどのようなことなのでしょうか。

山西氏：大量データの活用というとき、これまでは主に、統計学で言う大数の法則と中心極限定理で記述されるような対象を考えていました。つまり、データは基本的に、定常で一様な性質を持つ要素の集まりであると想定され、データ数が増えれば、平均値やその周りのばらつきさえ考えれば何が起こっているか把握できるような世界です。

現在のビッグデータは、容量（Volume）、スピード（Velocity）、種類（Variety）の3Vで定義されています。最近では、正確さ（Veracity）や価値（Value）を加えて、5Vと定義されることもあります。ビッグデータでは、容量（Volume）に注目が集まりがちです。しかし、私は次世代のビッグデータ分析ではスピード（Velocity）と種類（Variety）がより重要になってくると思います。例えば、大量のデータが蓄積されつつある医療の世界では、一人の患者に限ってみれば実はデータ量はそれほど多いわけではなく、大数の法則が成り立つ世界ではありません。しかし、その背後には同様の症状や症例のデータが大量に存在しています。その大量データを関連させることで浮かび上がってくる情報の中から、

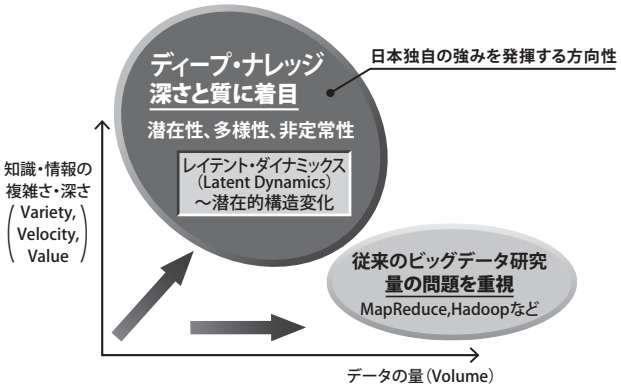

図3. ビッグデータ研究におけるディープ・ナレッジの位置付け

最適な治療方法を選択、決定できる可能性があります。

このように多様であり、広がりがあり、時間的にも、空間的にも変化しているデータを有効活用することがデータ・マイニングの本質です。こうした本質を考えていくと、現在のデータ・マイニングは、従来のような「定常で静的」「同質で一様」な性質だけでは説明し尽くせないところにきています。次世代のデータ・マイニングでは動的でヘテロな関係ネットワークを構成するデータの分析が鍵であり、それによって発掘すべき対象こそが「ディープ・ナレッジ（Deep Knowledge）：深層知識」なのです（図3）。ディープ・ナレッジは、データの表層からでは見えない、潜在的な関係性や、その構造変化を含めた深い知識のことを意味します。これは先に述べた「レイテント・ダイナミックス」を含む、より広い概念です。

企業の経営者や分析担当者にとって、ディープ・ナレッジの特性を生かし、少量、大量にかかわらず、手元にあるデータと世の中に存在する膨大なデータを横串に刺して、そのデータの全体像をいち早く俯瞰することが次世代データ活用の最大のポイントになります。データは企業の資産です。しかし一社で抱え込むのではなく、機密情報や個人情報などを守りながら公開可能な範囲で積極的に外部に公開し、例えば株価や気象、交通、通信などの情報と関連付けて、データが持つ潜在的な構造やつながりを制することが重要です。「情報コンソーシアム」のような仕組みを確立するのも有用です。データの価値を社会全体で共有し、最大化していくということです。

機械学習の分野でも、データ共有のためのプラットフォームづくりに大きな関心が寄せられています。別の言い方をすれば、ヘテロな要素がつながり合ったネットワークという視点の重要性が認識され始めています。

ここで言うネットワークでは、「どのような階層的つながりに

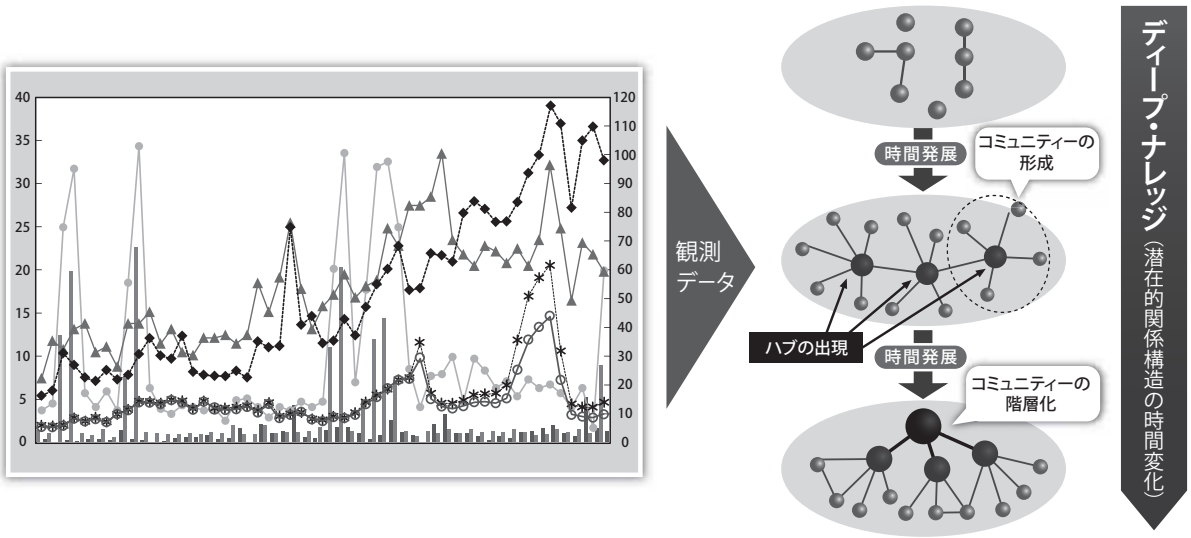

図 4. ディープ・ナレッジの例：ネットワーク・コミュニティーの構造変化

なっているか」「どのようなサブグループが形成されているか」など、ネットワークに潜在的に存在する構造を把握することがデータの本質を見いだすための鍵になります。これによって、ネットワークを構成する要素のランキングなどもできるようになります。例えば、スポーツ選手の属性や対戦成績が一部しかなくても、広大な関係性のもとで、戦ったことのない選手同士の間にもランク付けやグループ化を行うことができるようになります。しかもネットワークは時間とともに変化します。テキスト・データや数値データ、言語データなどがつながりながら、空間的、時間的に変動している潜在的世界の構造や動きを、客観的データに基づいて科学的な思考で解明することが、現在期待されていることなのです（**図 4**）。

これまでのビッグデータ分析は、相関係数の計算やクラスタリングなど、データを集計することで得られる知見をすくい上げることに重きが置かれていました。しかし、ディープ・ナレッジが扱う世界はより深く、潜在的な構造や因果関係のつながりなど、ネットワークの変化を捉えたものと言うこともできるでしょう。

診断で終わるのではなく 対処という新たな価値を創造

―― 企業におけるデータ活用の課題についてどのようにお考えでしょうか。

山西氏：データが価値に結び付いていない、次の意思決定につながっていないことが多いのではないでしょうか。データを価値へ結び付けるには、機械学習やデータ・マイニングで、どのようなデータを抽出するかを検討することに始まり、適切な加工と可視化を経て、価値を引き出すというステップに進みます。しかし「少しのデータから大きな価値を引き出して、アクションへつなげる」という最もスマートな方法で解決されず、いまだに直感的で属人的な暗黙知にとどまっている傾向があります。これを形式知化する努力が今後はますます重要になるだろうと考えています。

例えば、経営者がデータから得られた知見をどのように活用したのか、どう使えば有効だったのかという事例や経験を体系化し共有することで、第 2 のデータ・マイニングというべき、価値を引き出すサイエンスが実現されます。しかし、例えば異常検知において異常値が見つかった場合も、その現象に対処するだけで済ませてしまうと、「発見された異常値はどのような事象またはトラブルへの予兆であり、その異常値が検出された段階ではどのような対処法が有効であるか」といった知識には結び付きません。体系化することで、得られた情報は知識となります。知識はまた、異常時の的確な対応や予防的検知の開発という価値につなげることが大切です。異常検知から知識、価値にいたるプロセスの体系化を行い、事例を蓄積していくことで、データから価値を引き出すことができるでしょう。

そこでこれから活躍が期待されているのが、データ・サイエンティストという新しい職業です。データ・サイエンティストは、データ活用における医師とも言えます。医療現場では医師が患者の診断結果や症状をモニタリングして、異常があればアラートを出し、原因は何か、合併症はないかなどを総合的に判断して治療という価値を提供します。これと同じように、データ・サイエンティストは企業の活動をモニタリングして、異常発

生時にはその原因を特定し、ほかのビジネスとの関連を調査して、いかに対処すれば企業が健康になるかを提案します。

以前、ある海外の大企業の不正経理事件が世の中を騒がせたことがありますが、その事件の裏でやり取りされていた大量の社内メールを分析すると、経営が傾いていくときに発生する異常のパターンを特定できることが分かりました。これはまさにディープ・ナレッジです。このように企業内で飛び交うメールを分析するだけでも、企業の健康状態を診断できます。またこのようなデータは、自社のデータを囲い込むのではなく、多様なデータと関連付けて大きなネットワークで見ていくと、さまざまな発見があります。こうした「情報診断」をやりながら、次のアクションを意思決定して実行し、価値を生み出すことが重要なのです。

いわゆる分析ツールを使って現状の表面的な統計量を把握するだけで満足する人が多いのも事実です。しかし真に有意義な知識はディープ・ナレッジを分析することで得られると思います。データ間の相関ルールを見つけておしまいではなく、その構造や因果関係、時間的変化など、一歩深入りした潜在世界に目を向けて価値を引き出すことで初めて「ディープ・ナレッジ」へと到達できます。そうして獲得されたディープ・ナレッジを束ね、ネットワーク化して企業に価値を還元していくことこそ、データ・サイエンティストの役目であると考えます。

データ・サイエンティストが照らす データ活用の未来像

―― 将来的な展望について聞かせてください。

山西氏：ディープ・ナレッジの世界では、いろいろなデータがつながる中で形作られる潜在構造に、実世界で有用な知見が埋め込まれていると考えます。そこでは、データの規模だけに着目した解析手法ではなく、潜在構造の多様な変化から最も適切なモデルを同定し、その変化を捉える方法論が重要です。繰り返しになりますが、大量のデータがあっても、ヘテロな要素の一つに限ってみれば、対象となるデータは多くはありません。そのような場合でも、少数データの背後に広がっている膨大なデータとのつながりの構造に基づいて、データを効果的に連携し、価値を見いだすことが重要です。そのような、多種多様なデータが関連付けられて広がりを持つ世界に光を与えるのがデータ・サイエンスの世界です。

データ・サイエンスの今後の発展のためには、私は、データ・サイエンティストの社会的な価値を向上させていくことが大切だと考えています。

医師や弁護士になるためには資格が必要です。資格が社会的地位に直結しています。一方、データ・サイエンティストは“世界で最もセクシーな職業”と言われ、米国で優遇されつつある職業ではありますが、資格ではなく、日本ではその社会的地位はまだ確立されていません。しかし、私は将来、データ・サイエンティストも高い社会的地位を獲得できるようになるだろうと考えています。その際に、データ・サイエンティストの試金石は、ディープ・ナレッジを発掘できる能力があるかどうかであると考えます。

診断するだけでなく、治療によって価値を創造することが必要であるという話をしましたが、データ・サイエンティストは、例えば解析ツールを使って相関ルールを見つけたり回帰分析をしたりできるだけでなく、それらが「なぜ」「どのように」つながっているのかを俯瞰して見ることができ、それを実世界の価値へ結び付ける能力が必要です。そこまで実現できて、初めて「ディープ・ナレッジを活用できている」と言えるでしょう。そこからさらにビジネス上の真の価値につなげるには、データ・サイエンティストだけでは不十分で、企業のビジネス・ドメインを熟知している経営者や現場の担当者とのコミュニケーションが欠かせません。分析の専門家とデータを持つ企業が組むことで、目標である価値創造へ到達できるはずです。わが国においても、そのような能力を持ったデータ・サイエンティストがたくさん登場し、高い社会的地位を獲得する日が来るのもそう遠くないと思っています。