特集「発見科学」にあたって

元田　浩
（大阪大学産業科学研究所知能システム科学研究部門）

「発見」という知的行為は常に人々を魅了し，あらゆる分野で人類の進歩の原動力になってきた．しかし，哲学の長い歴史の中でも「発見」は議論の対象から外れており，１９６０年代になり新科学哲学派によってはじめて本格的な議論が開始されるに至った [Noe 1998]．人工知能の研究分野では，ここ２０年の間に経験的発見という名の下に発見の機械化の研究が地道に進められてきた [Langley 1989].「発見」と「学習」が持つ意味合いは違うが，歴史的には発見は観測からの学習の難しい形態であると見なされている [Langley 1986]．一般的には，学習は緩やかなプロセスであるが，発見は洞察を伴うもっと迅速なプロセスである．学習は眠っていても進む（無意識に知識が変化する）が，発見は常に意識下で起こる．学習の結果得られる知識は宣言的であることも手続き的であることもあるが，発見の場合は常に宣言的である．このような違いがあるにもかかわらず，発見の機械化の研究で使われている手法の多くは機械学習の成果である．発見も学習も「知識獲得」の一種である．しかし，知識工学の分野では，知識獲得は人間の専門家から知識を獲得し計算機で使用できる形に変換することを意味しており，意味合いが違う．一方，ディジタル時代に入り，我々の処理能力を遥かに上まわる多量の計算機可読なデータの洪水の中に溺れてしまう現象があちこちで生じ始め，データの中から有意な知識を発掘するデータマイニングのニーズが急速に高まってきている．これには機械学習の技術はもちろん，データの可視化技術，長い伝統のある統計やパタン認識の技術が統合して使われている．これらを鑑みるに，従来，どちらかと言えば個別に進められてきた「発見」に関する研究を総合して「科学」する土台ができ,「発見科学」という新しい学問を創設する機が熟したと言っても過言ではないであろう．

このような背景から「発見科学」に関する特集を組むに至った次第であるが，発見に関連する分野は多岐にわたり，それらを個別に解説したのでは，単に個別の分野の解説記事の寄せ集めに終わってしまい，発見科学の特徴が出せない恐れがある．まだ細部を解説できるほど整理されていない新しい分野であるので，本特集では解説論文は１編に留め，残りを研究論文とし投稿を募集することにした．平成１０年度に発足した文部省科学研究費補助金特定領域研究（Ａ）「巨大学術社会情報からの知識発見に関する基礎研究（略称：発見科学）」[有川 1999, 有川 2000] も２年目が終了したところであり，総括責任者：有川節夫教授（九州大学）と５名の班長：佐藤雅彦教授（京都大学），佐藤泰介教授（東京工業大学），丸岡章教授（東北大学），宮野悟教授（東京大学），金田康正教授（東京大学）に発見科学プロジェクトの目標，成果などを総合的に紹介して頂き,「発見科学」全般の解説に当てさせて頂くことにした．このプロジェクトは発見科学を「知識発見の論理」,「推論による知識発見」,「計算学習理論に基づく知識発見」,「巨大データーベースからの知識発見」,「ネットワーク環境における知識発見」の５つの基本課題（班）に整理し，各班長の指揮の下に総勢数十名の研究者が協力しあいながら，それぞれの基本課題に挑戦し，知識発見を１つの学問分野として自己発展可能なレベルまで高めることを目指している．

研究論文に関しては，２０件以上の応募があり，発見科学への関心の高さが伺えた．学会誌の紙面の制約から本特集号には１０件の論文しか採録できなかった．幾つかの論文は一般論文に回さざるを得なかったことをご了承頂きたい．各論文とも研究成果だけでなく，発見科学に関する研究の位置づけ，同分野の研究の現状も書いていただくようにお願いしたので，それぞれの研究の周辺状況もある程度はお分かり頂けるものと思う．以下，１０件の論文を簡単に紹介する．

1. 科学者の類推による発見

人間の認知的活動の観測を通して科学的発見に至る仕組の手掛かりを得ようとする試みは多いが，従来の研究は歴史的文献や科学者の逸話の分析や実験室環境での問題解決の分析に留まっていた．具体的な研究活動の詳細な分析を通して認知活動を解明しようとする研究はつい最近始まったばかりである．植田論文は科学的発見を可能にする認知メカニズムを類推と関連づけて詳細に分析したものである．物質科学，生化学，生態学，天文学の分野の国内の２２人の著名な研究者にインタビューし４２の事例を収集し，それを詳細に分析し半数近くで類推が頻繁に利用されていることを確認している．さらに，類推を３つの類似性基準と２つの転写原理からなる６つの組合わせに分類し，観察した事例がこのうちの４つの組合わせに該当し，理論構築などの発見に近いものは新た

なカテゴリー形成に基づく類推と因果構造に基づく転写の組合わせであることを突き止めている．インタビューという非常に手間のかかる作業を通して認知科学的に貴重なデータを提供している論文である．

2. テキストデータからの高速データマイニング

計算機可読な文書が多数蓄積されるにつれ，テキストデータベースと呼ばれる新らしい形のデータベースの利用がここ数年急速に進んでいる．テキストは明示的な構造を持たず，非均質かつ量が膨大であるため，通常のトランザクションデータを念頭に開発されたマイニング手法や関係データベースを対象に開発されてきた検索手法は使えない．阿部・藤野・下薗・有村・有川論文は彼らが開発してきた，多量の文書中に任意長の d 個の指定した文字列が互いに k 単語以下の距離をあけて指定の順序で連続して出現（近接語相関パターン）する頻度を一般化接尾辞木を利用して高速に算出する方法を基に，正例の文書と負例の文書をもっともよく分離する近接語相関パターンを発見する手法を提案したものである．探索的文書ブラウジングと WWW ページからのキーワード獲得に適用し非常によい結果を得ている．統計的決定理論や計算論的学習理論とも理論的基盤を共有する深みのある研究であり，かつ実装面でも多くの工夫をして実用性を高めることに成功した研究である．

3. データベースからの知識発見システム DB-Amp

表現力の高い一階述語論理に基づく帰納論理プログラム (ILP) は，既存知識を背景知識に使いデータから新しい知識を発見し，それをまた背景知識に加えることが出来る有望な帰納学習の方法であるが，既存のデータベースのデータを直接使うことはできなかった．嶋津・古川論文はデータベースと ILP を統合する手法に関するもので，データベースから ILP の入力データ（背景知識，正事例，負事例）を自動生成し，ILP にて新たな知識を学習するシステムを提案している．電子メール応答記録データベースのデータを用い，問合わせ文と回答文を関連づけるルールを学習し，新たな問合わせに対し，推奨される回答文を表示するシステムを試作し，実際にオペレータの質問回答業務時間が削減されたことを報告している．まだ，ILP の入力データ生成に時間がかかっているが，ILP の実用化への１つのステップを示したものと言えよう．

4. 適切な抽象化に基づくデータベースの一般化によるデータマイニング

相関ルール解析をしてみれば分かるが，ルール発見では極めて多数（数千以上）のルールが抽出される．そのほとんどが既知のものであったり，面白くないものであることが多い．興味深いルールだけを選択的に発見することが重要な課題となっている．工藤・原口論文はデータベースの属性の値には視点の違う複数の階層化の方法があることに着目し，抽象化によっても情報利得が変わらない視点と階層レベルを見いだしデータベースを一般化することにより良質の知識を発掘する手法と，抽象化によって同一グループにまとめられる事例の規模がある程度以上のものを無視することにより隠れた特徴を見いだす手法を提案している．これにより，分類精度を落とさない簡単な決定木の生成や一般的な特徴の裏に隠されていた面白い特徴の発見が可能となることを示している．

5. 閾値スケジューリングに基づく仮説駆動型例外ルール発見

鈴木論文も上記の興味深いルールだけを発見する問題に対して別の視点からアプローチしている．常識的ルール（これは通常の相関解析で容易に発見されるルール）とそれに幾つかの条件を加えたら結論が変わる（加えた条件単独ではそのような結論は示唆されない）例外的ルールのペアを同時に求め，後者を興味深いルールとして，その数を自由に制御できるよう探索の閾値を自動的に制御する手法を提案している．探索アルゴリズムは高速であっても入力の閾値をどう設定すればよいかという問題が残されていたが，これが解決され例外ルールの自動発見に近づいた．

6. 論理最小化に基づく決定木による知識発見

属性選択手法の多くは C4.5 に代表されるように各属性とクラスの相関の強さのみを選択基準とするものが多く，属性間に強い従属関係がある場合はうまく行かないことが知られている．稲葉・吉澤論文はこの問題に対する１つの解決法を提案している．論理最小化手法 MINI における各論理変数の論理式簡略化への貢献度が各属性のクラス分類に対する貢献度と対応づけられることに着目し，貢献度を定量的に評価する選択基準を提案し，C4.5 では上手く解けない問題を解決している．まだ小規模のデータセットでの結果しかなく，連続属性，離散多値属性，ノイズを含むデータに対して問題を含んでいるが，今後，大規模問題に適用し良好な結果が得られれば，実用性の観点からも有望な方式になると思われる．

7. キーワード抽出法 KeyGraph の転用による地震履歴データからの要注意断層発見支援

大澤・谷内田論文は著者らが開発した文章からキーワードを抽出する方法を地震の発生予測に応用した面白い論文である．地震の発生の履歴データから発生個所にもっとも近い活断層を探し，それを単語と見なし単語の系列を得る．これを文と見做し，大きな地震が起きた直後を文の切れ目とし，文章を作成する．文章には土台があり，その上に柱が立ち，それに支えられた屋根（主張）があるとの発想で，単語の頻度と共起関係から，これら３つを同定し主張を抽出する．これがキーワードであり，地

震データに対しては要注意断層になる．ある時期以前に発生した地震データから，実際にそれ以降に発生した地震を見事に予測している．さらに，活断層の移動や今後地震の発生しそうな場所を予測している．地震学者の知見とも一致しており，有効な地震予測支援システムになると思われる．

8. 時系列モデルによる大量データからの情報抽出

近年の計測技術の飛躍的な発展により大量に観測され蓄積される時系列データを解析し，その中から有益な知識を獲得したいと言うニーズはますます増加している．大量の継続的な観測時系列データを解析する手法は古くから「時系列解析」として数理統計学の一大分野を形成している．北川・松本論文は，データ数に比例する処理時間で，欠損値と異常値をフィルタリングし対象とノイズを分離する状態空間モデルを用いた解析法が，地震による地下水データの解析を例に，非常に微妙なデータの変化を正しく検知できることを示している．パラメータ同定を中心とする従来の統計学とちがい，積極的に対象に関する知識や期待を埋め込むことが可能な状態空間モデルの威力とこの分野のレベルの高さを示す論文である．

9. スケールタイプ制約に基づく科学的法則発見式の発見

古来，物理学者が実験データを深く分析し裏に潜む真理を見出してきたように，人間には手におえない多量の実験データ（数値）からデータ間の関係を支配する第一原理法則を計算機を用いて機械的に発見することができれば，実用上の効果は大きく，かつ「発見科学」にも貢献する．鷲尾・元田論文はその可能性を模索したものである．できるだけ対象領域の知識を使わないで，実測値の測度の性質の違いを制約として，測定データ間に許される関係式を一般的に求め，これに基づき表現された関係式の範囲の中から実際に得られた観測データを説明できる具体的な関係式を実験的に求める手法を提案し，数値実験により既知の法則の再発見や 10 数変数の回路方程式の同定に成功している．演繹的な次元解析の手法とデータからの帰納推論手法を組合わせたもので，得られる関係式の完全性をある程度保証している点に特徴がある．

10. WWW 情報の構造視覚化と検索機能の統合

WWW は情報の倉庫であるが，あまりにも多量の，しかもよく構造化されていない情報の集りであるため，その中から有益な情報を迅速に発掘することは非常に難しい．多くの検索エンジンが提供されているが，適切なキーワードの指定が難しいこと，多量の本題に関係のない検索結果が戻されてくるなどの問題点が指摘されている．そのため WWW ブラウジングを支援する研究は非常に盛んである．大和田・溝口論文は情報の視覚化と検索の機能を統合したインテラクティブな情報検索支援システムに関するもので，一般ユーザを被験者に，所望のサイトの所望のテキストに辿りつくまでのマウスのクリック数や訪問したサイト数を測定し，実際に効果があることを報告している．焦点を自由に移動させながら WWW の全体構造を立体的に表示する方法や文書間の類似度を評価した検索法はユーザに使って見たいという気持ちを起こさせる．

最後に本特集号にご協力頂いた各執筆者の方々に感謝の意を表します．

◇参　考　文　献◇

[有川 1999]　有川節夫（領域代表者）：発見科学－巨大学術情報社会からの知識発見に関する基礎研究－, (1999)

[有川 2000]　有川節夫（領域代表者）：発見科学－巨大学術情報社会からの知識発見に関する基礎研究－, (2000)

[Langley 1986]　Langley.P., “Editorial: Machine Learning and Discovery,” *Machine Learning*, 1, pp.363–366, (1986)

[Langley 1989]　Langley.P., “Data-Driven Approaches to Empirical Discovery,” *Artificial Intelligence*, 40, pp.283–312, (1989)

[Noe 1998]　Noe, K., “Philosophical Aspect of Scientific Discovery: A Historical Survey,” *First International Conference, DS'98, Lecture Notes in Artificial Intelligence*, LNAI 1532, Springer-Verlag, pp.1–11, (1998)