■S3 群（脳・知能・人間）－ 3 編（人工知能と学習）

8 章 知識ベース応用システム

（章主任：山口高平）［2018 年 11 月 受領］

本章では，知識ベース応用システムとして，エキスパートシステム，知的検索システム，ナレッジマネジメント，社会インフラオペレーション，知的教育支援システム，バイオインフォマティクスを取り上げ，知識ベース及びその他の人工知能技術の適用方法について説明する．

エキスパートシステムでは，知識ベースと推論エンジンを分離した問題解決システムであり，開発初期段階で頻繁に知識が修正される場合，知識ベースと推論エンジンが一体的にコーディングされた通常のプログラミングシステムと比較して，開発が容易になることが大きな特色であり，世界初のエキスパートシステム MYCIN を例にとり説明する．

知的検索システムは，Web 検索システムと知識処理を統合したシステムであり，インデックス処理とサーチ処理における知的処理，及び，情報検索エキスパートシステムについて説明する．

ナレッジマネジメントは，知識の収集，蓄積，更新，分配，共有，創出などを含む一連の知識操作プロセスであり，検索エンジンやワークフローなどの一般的な情報サービスとの関連もあるが，本節では，知識表現とデータマイニングの適用事例をもとに，高度なナレッジマネジメントについて説明する．

社会インフラオペレーションでは，状態監視，結果の判断，アクションまでを含むオペレーションの最適化を目指して，結果の説明が容易な機械学習技術が採用されており，上水道オペレーションと原子力発電所異常予兆検知の事例について説明する．

知的教育支援システムでは，学習者の誤りの修正を中心とした知識伝達に取り組んだ ITS（Intelligent Tutoring Systems），学習者自らの知識構築を支援する学習環境 ILE（Interactive or Intelligent Learning Environments），学習者間の協調的知識共有や知識構築を支援する CSCL（Computer-Supported Collaborative Learning）を中心に説明する．

バイオインフォマティクスでは，代表的な分子生物学データベースを紹介し，類似配列検索，配列からの機能予測に関する，機械学習を中心とした人工知能技術の応用について説明する．

【本章の構成】

本章では以下について解説する．

■S3 群－3 編－8 章

8-1 エキスパートシステム

（執筆者：山口高平）［2018 年 11 月 受領］

エキスパートシステム（Eexpert Systems，以下 ES と略記）とは，専門家の持つ専門知識をコンピュータ内部で表現し利用することにより，専門家のように知的に振る舞うシステムである．1970 年代中期，スタンフォード大学において，感染症の原因となる病原菌を同定し，薬の投与法を決定する，感染症診断 ES MYCIN（マイシン）が世界初の ES として開発された．

ES の構造を**図 1･1** に示す．ES の最大の特色は，知識ベースと推論エンジンが分離されていることである．問題解決に必要な専門知識とその利用方法をプログラム中に混在して表現すると，経験的に得られた専門知識は，内容が確定するまで頻繁に修正されることが多く，一つの知識の修正がプログラムレベルでは修正が散在し，修正コストが大きくなるため（1970 年代，民間会社からそのような報告がされている），専門知識をシステム化する場合，知識ベースと推論エンジンの分離が鍵となった．

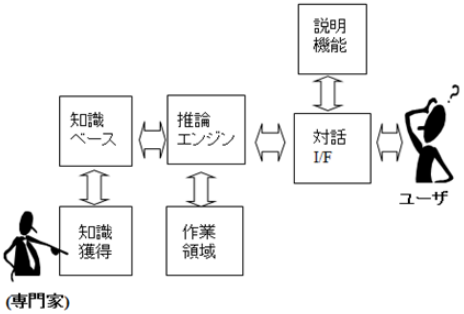

図 1・1　エキスパートシステムの構成

ES の知識ベースの多くは，IF（条件部）-THEN（結論部）というルール形式で表現され，ルールの利用方法（推論）としては，IF から THEN に処理が流れていく前向き推論と，THEN から IF に処理が流れていく後向き推論がある．上述した MYCIN は，500 程度のルールから成るルールベースを持ち，LISP 言語により，後向き推論が実装された．図 1･1 において，前向き推論では，観測データが作業領域に保存され，推論エンジンが作業領域とルール条件部を照合し（照合），複数のルールの条件部が照合されれば，競合解消戦略（例えば，より多くの条件節を持つルールを優先するなど）により一つのルールを選出し（競合解消），選出されたルールの結論部を実行する（実行）という，認知-実行サイクルを結論が得られるまで，繰り返し実行する．一方，後向き推論では，成立すると仮定した仮説を結論部に持つルールから，そのルールの条件部にある条件節に展開し，更に，その条件節を結論部に持つ別のルールがあれば，その別のルールの結論部から条件部に展開し，最終的に展開不能となった条件節に対して，それが成立するか否かをユーザに確認し，一つでも不成立ならば，そのルールを棄却し，別の仮説を結論部に持つルールに移行して，同様の操作を繰り返し，展開されたすべての条件節が成立する場合，その元ルールの結論部を推論結果とする．

なお，ユーザが条件節の成立可能性について尋ねられたとき，ユーザは（Yes 確信度 0-1.0）を入力し，ルールの確信度とユーザ入力の確信度を考慮して，結論の確信度を計算する仕組みも考案され，これを不確実性推論と呼ぶ．

また，ES は，ユーザが ES の振舞い（ユーザに提示する質問や結論）に疑問を持ったとき，ユーザが Why と How コマンドにより説明機能を利用できる．**図 1･2** は，ユーザである医師と MYCIN の対話例を示すが，医師は，MYCIN からの質問 38)サリーは，やけどをしていますか？という質問をした背景理由が理解できなかったため，Why（なぜ今の振る舞い＝38）のような質問をするのか？）と入力して説明を求めたところ，MYCIN はその質問提示に関与した条件節を持つルールを提示し，現在，結論節（感染症の原因は緑膿菌である）を仮定したルールを処理しており，第 4 条件節の成否を確認するため，38)の質問をしたと回答している．また，このルールの第 3 条件節は，MYCIN が自動的に成立したと判断しており，医師がその理由を尋ねるために，How(3)（条件節(3)をどのように処理したのか？）と入力したところ，ルール 500 と 501 を使って処理したと回答している．

```
37) サリーの臨床歴に中耳炎がありますか？
はい.
38) サリーは，やけどをしていますか？
Why

現在，下記のルール384を使って診断中ですので，★(4)の条件節を質問しました.
ルール384
IF      ☆(1) 治療を必要とする感染症は髄膜炎である.
        ☆(2) ＣＳＦ培養物の標本に細菌は含まれていない.
        ☆(3) 感染症の型は細菌によるものである.
        ★(4) 患者は重度のやけどをしている.
THEN  (1) 感染症を引き起こしている原因菌のひとつは，緑膿菌である

Ｈｏｗ（３）
（どのようにして，感染症の型が細菌によるものであると判ったのか？）

以下のルールにより，感染症の型が細菌によるものであると結論付けた.

ルール５００     (Yes, ０．４)
ルール５０１     (Yes, ０．６)
```

図 1・2　医師と MYCIN の対話

Edward Feigenbaum 教授は，IJCAI（人工知能国際会議）1977 において，MYCIN の成功例について言及し，知識を獲得し，コンピュータ内部で表現し，利用する研究分野を知識工学（Knowledge Engineering）と命名し，AI の一分野を形成することになり，ES に対する産業界・社会の関心が高くなっていった．こうして 1980 年代以降，コンピュータ，鉄鋼，建設，電力，石油，化学，機械，ビジネスなど様々な産業界で，診断，スケジューリング，設計支援などの ES が全世界で 5000 程度開発され，ES が第 2 次 AI ブームの牽引者となった．

しかしながら，知識分析の専門家（知識エンジニアと呼ばれた）が領域専門家にインタビューして，ルールベースを構築していくことは，大変時間のかかる作業であり，ES の開発には知識獲得ボトルネックがあると批判された．更に，領域専門家が別の部署に異動し，ルールベー

スの開発背景や文脈は外在化されていないので，新しく異動してきた領域専門家にとっては，ルールベースの意味を把握できず，ルールベースの維持発展が困難となり，知識獲得ボトルネックに知識維持ボトルネックも加わり，1990 年代以降，ES はほとんど開発されなくなった．

この課題を受けて，ルールベースや推論エンジンに仕様を与えるべきという議論が始まり，モデリングプリミティブの仕様としてのオントロジーの研究が開始され，セマンティック Web や LOD（Linked Open Data）研究が現在進んでいる．また，ES が，自然言語入力 IF とプログラム自動出力 IF を備えて，ビジネスルール管理システム（BRMS: Business Rule Management Systems）に変貌し，アジャイルな業務システムの開発及び保守を実現し，ビジネスツールとして認められていることは興味深い．

■参考文献

1) 山口高平（分担翻訳）：“人工知能における知識ベースシステム,”啓学出版, l99l.
2) 森田武史, 山口高平：“業務ルール管理システム BRMS の現状と動向,”人工知能学会誌, vol.29, no.3, pp.277-285, 2014.5.

■S3 群－3 編－8 章

8-2 知的検索システム

（執筆者：松田勝志）［2016 年 7 月 受領］

8-2-1 知的検索システムの概要

1990 年代前半までの情報検索技術では，司書や弁理士など図書館業務や特許業務を行う専門家を主な利用者と想定していた．しかし，1990 年代後半から Web の普及が本格化し，Web 検索やオンライン蔵書目録（OPAC）や特許電子図書館（IPDL）などが Web 上で利用可能となることで，非専門家であるインターネットユーザが利用者となった．情報検索技術の評価検証を行うプロジェクトである TREC（Text Retrieval Conference）や NTCIR（NII Testbeds and Community for Information access Research）でも 1999 年と 2001 年にそれぞれ Web トラックが始まっており [1),2)]，Web 検索が重要な検索技術であるとの認識が広がっている．特に，利用者の多い Web 検索サービスのコモディティ化は，専門家のための検索から非専門家のための検索へと対象ターゲットの大変換を起こした．非専門家が専門家と同等の検索速度や検索精度を達成するためには，検索システムによる専門家のスキルに相当する知的な支援が必要である．非専門家を支援する知的な処理を組み込んだ検索システムを知的検索システムと呼ぶ．

図 2･1 にハードウェアを省略した Web 検索システムの一般的な構成を示す[*1]．検索システムは，収集エンジンがインターネットから Web 文書を収集し文書 DB へ登録，索引エンジンがインデックスを作成，ランキングエンジンが文書のスコアを計算してインデックスに登録するまでのインデックス処理と，利用者がクライアントから検索アプリを通じて検索要求を行い，検索エンジンがインデックスを用いて検索結果を検索アプリに返却，検索アプリが検索結果をクライアントに表示するサーチ処理に大別できる．

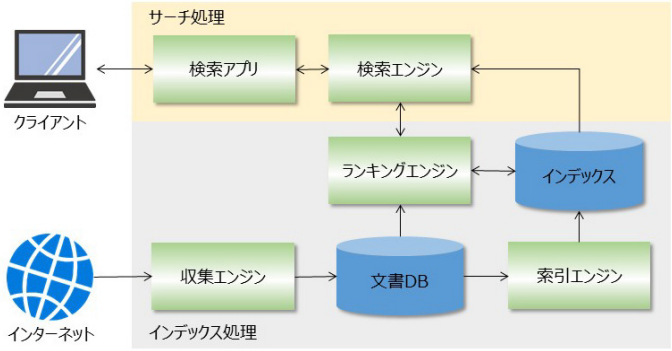

図 2・1　Web 検索システムの構成

知的検索システムは，図 2･1 の Web 検索システムを構成する 1 つ以上のコンポーネントに

*1　ここでは Web 検索システムの構成を示したが，図書検索システムや特許検索システムもほぼ同様な構成である．図書検索システムならば収集エンジンは他図書館の OPAC から文書を収集，特許検索システムならば収集エンジンは公開特許情報から文書を収集することになる．

知識処理を組み込んだものである．厳密には明確に区別できないが，サーチ処理では専門家と同等の検索速度，インデックス処理では専門家と同等の検索精度を実現することを目的とした技術開発や研究が行われている．検索速度とは，利用者がすばやくクエリ（検索キーワード）を作成，入力，検索結果をすばやく俯瞰，するなどの指標であり，検索精度とは，検索対象を大幅に絞り込んだり（適合率を上げる），検索漏れを減らしたり（再現率を上げる），求める文書をより結果上位にするなどの指標である．検索精度の評価尺度である適合率（Precision）と再現率（Recall）についてはここでは述べない．評価尺度については文献 3）や本知識ベースの 7 群 6 編が詳しい．以下，インデックス処理とサーチ処理それぞれにおける知的化技術の先駆例を紹介する．

8-2-2 インデックス処理における知的化

Web 検索システムにおける収集エンジンの主な機能は，検索対象の文書を大規模かつ効率的に収集することである．ここでは大規模化について注目する．検索対象の文書数が増えると，求める情報を含む文書（適合文書と呼ぶ）が検索結果に出現する個数は多くなる一方，それ以上に求める情報を含まない文書（不適合文書と呼ぶ）が検索結果に出現してしまい非専門家にとっては適合文書を見つけることが難しくなる．現在主流の n-gram[4] を用いた転置インデックス [5] 方式の全文検索エンジンの場合，クエリとして与えた検索キーワードは検索結果すべての文書に含まれる．検索キーワードは求める情報を端的に表現したものではあるが，検索キーワードが含まれているから求める情報というわけではない．

求める情報の種類を大まかに分類（ページタイプ）し，その分類に含まれる文書を集中的に収集することで検索結果の不適合文書の数を減らす試みがある [6]．ページタイプには，商品カタログ，オンラインショップ，求人案内，論文募集，リンク集，FAQ，用語集，サイトトップページ，掲示板などがある．各ページタイプは，Web 文書に含まれているキーワード，各種タグ，URL の文字列，画像の有無などの構造的特徴により分類する．収集エンジンがページタイプに分類しながら Web 文書を収集することで，目的別の検索システム，例えば求人案内のみを検索対象とする求人案内検索システム，を提供することができる．また，同技術を収集エンジンの収集先判断に組み込むことで，特定ページタイプの Web 文書のみを集中的に収集する事例も紹介されている [7]．

収集エンジンにおいて不適合文書を削除するその他の技術には，Web 文書やコンテンツ部分を不適合と人が判断する行為をモデル化した重複文書検出技術 [8] やエントロピーに基づくブロック判定技術 [9] などがある．

全文検索の場合の索引エンジンは，転置インデックスでほぼ解決をみており [10]，この部分で知的な処理を行う余地は少ない．内容型検索 [3] の場合は，LSH（Locality-Sensitive Hashing）[11] という類似検索に適したハッシュ法が考案されて以降，研究が活発になっている．

Web 検索システムにおけるランキングエンジンは，検索結果から適合文書を特定するための指標を生成する機能を担っている*2．Web 文書の求める情報への適合度をスコアとして表現す

*2 図 2･1 の検索システムの構成では，ランキングエンジンはインデックス処理において実施されることになっているが，ランキングの手法によっては，サーチ処理，すなわち検索エンジンが検索結果を返却するときに動的にスコアを付与することもある．例えば，tf・idf [12] の場合は，検索キーワードの出現頻度をスコアの計算に使うため，事前に計算しておくことが困難である．

る．検索エンジンがスコア順に検索結果を返却し，利用者は検索結果の上位から参照して適合文書を探す．利用者が検索結果を見るのは上位 10～30 件である [13),14)] という報告もあるように，Web 検索システムは数万件から数百万件に及ぶ検索結果の上位 30 件，可能ならば上位 10 件に利用者の求める情報を含む Web 文書を表示することが求められている．すなわち，検索結果から求める情報に到達するまでの速度を上げるための技術である．検索キーワードから利用者が求める情報を推定することは非常に困難であるため，ランキングエンジンは一般的に何らかの検索意図モデルに基づいてスコアリングを行う．

Web 検索システムで優位性が認められている検索意図モデルは，様々な Web 文書からリンクを張られている重要な Web 文書を探すことが多いというものである．代表的な技術は PageRank [15)] や HITS [16)] である．PageRank は，Web 文書をノード，リンクをエッジとした有向グラフとみなし，その推移確率行列の固有ベクトルを求めているが，公開された数式のみに基づいてスコアリングをすると，PageRank の値を故意に上げることが可能となるため，2009 年の時点で 200 以上のルールを用いてスコアの調整を行っている．具体的なルールは非公開だが，パラメータの一部は公開されている．例えば，IP アドレスの更新頻度，サーバの稼働率，正しい HTML が使われているか，タグとテキストの比率，内部リンクの数，サイトのページ数，ページの更新頻度などである．

8-2-3 サーチ処理における知的化

検索エンジンは，クエリに基づいてインデックスを検索し，検索結果及びそれらのスコアを返却する．検索アプリと検索エンジンの役割の明確な区分は困難であるが，ここでは利用者が入力した生の検索キーワード（または検索クエリ）をもとに検索エンジンが解釈できる検索式に変換するのが検索アプリで，検索式に基づいてインデックスを検索するのが検索エンジンとする．また，検索エンジンの返却する検索結果を利用者が閲覧するのに適した形式に変換するのも検索アプリの役割とする．

検索エンジンにおける知的処理には，あいまい検索，多言語検索，クエリ拡張などがある．あいまい検索（Approximate String Matching）は，検索キーワード文字列に対して，ある一定の範囲で挿入，削除，置換を行い，検索キーワードの誤字などの不具合を吸収することで検索漏れを減らす技術 [17)] である．多言語検索（Cross-Language Information Retrieval）は，入力された検索キーワードとは別の言語で作成された文書を検索する技術であり，検索キーワードを機械翻訳するキーワード翻訳型と検索対象の文書をあらかじめ機械翻訳してインデックス化しておくコンテンツ翻訳型がある [18)]．キーワード翻訳型は検索エンジンで，コンテンツ翻訳型は索引エンジンで処理される．いずれの場合もあいまい検索と同様に検索漏れを減らして再現率を向上させることを目的としている．キーワード翻訳型では多言語シソーラスや辞書を用いてキーワードをターゲットの言語に翻訳するが，キーワード数が少ないと正しい語に翻訳できないという問題がある．コンテンツ翻訳型は機械翻訳の品質や新語や新訳対応にコストがかかるという問題があり，検索エンジン単体での対応には限界があり，実用に耐えるレベルにはまだ達していないのが現状である．クエリ拡張は入力された検索キーワードに派生キーワードを付与することで検索漏れを減らすための技術であり，多言語検索と同様にシソーラスや辞書を用いて派生キーワードを生成する．1970 年代から 1990 年代には非常に活発な研究領域だったが，Web 文書の増大により検索漏れを減らすことに対する要求が減少しており，研究分野としては

活発と言えない．

利用者とのインタラクションを行う検索アプリは，Web 検索システムの知的化，すなわち，あたかも検索の専門家が仲介しているようにシステムが振る舞うという意味で最も重要なターゲットである．例えば，検索キーワードの入力を補助する音声入力やサジェスト*3 や位置情報検索，検索キーワードの入力を代替する適合フィードバックや類似文書検索，検索結果をリスト形式以外で表示する可視化技術などがある．ここではサジェストと適合フィードバックについて説明する．

サジェストは，Web 検索サービス大手の Google が 2008 年にリリースした機能であり，検索キーワードの入力時にインクリメンタルに検索キーワードの候補を表示し，選択可能とするものである [19]．膨大な検索ログと様々なパラメータを使って検索キーワードを予測している．インクリメンタルサーチ，すなわち検索キーワードの入力途中の振る舞いに関する研究分野はほとんどなかったが，アプリケーションソフトウェアの世界ではオートコンプリートとして例えば表計算ソフトなどで実用化されており，入力を支援する機能として有効であることが分かっている．検索アプリの機能としても，検索キーワードの入力を支援することで検索要求を開始するまでの時間を短縮する，記憶があいまいな語彙でも正しいものを入力して誤検索を防ぐ，すなわち検索エンジンにおけるあいまい検索と同等の効果がある．

適合フィードバック（Relevance Feedback）は，検索結果の中から利用者が求める情報を含む文書に近い文書を選び，その文書を元に検索キーワードを修正，または直接類似文書を検索する技術 [20] である．利用者は最初の検索キーワードのみを入力し，以降は類似文書を選択するだけでよいため，利用者の負担を減らす非常に有効な手段として様々な検索サービスで用いられたが，実際には想定ほど利用されなかったため，技術としてはある程度性能があるにも関わらず，実用には至らなかった．類似文書をチェックボックスで選択するというユーザインタフェースの問題と非専門家にはその操作によって何が得られるのか分かりにくかったという問題が原因と考えられる．プライバシーの問題などハードルが高いが，システムがバックグラウンドで検索操作や閲覧履歴を分析して自動的に検索結果に反映させるなどの工夫が必要であろう．

8-2-4 検索の知識ベース的アプローチ

情報検索全般における知識ベース的アプローチ，すなわち，非専門家を支援する仲介者としての専門家システム（Expert System）の研究は 1992 年以降大きな進展はない [21]．文献 21）の原著者は，情報検索エキスパートシステムに必要な機能を，領域モデル，システムモデル，利用者モデル，システムモデル適応器，利用者モデル構築器，検索戦略，応答生成器，フィードバック生成器，検索質問モデル構築器，写像，説明，変換器，計画立案者の 13 機能としている．これらのすべての機能を高度に実現するのは難しいため，検索システムの構成要素に対して知識ベース的アプローチを適用する方向に進んでいると考える．また，厳密には情報検索とは違うが知識ベース的アプローチは Watson [22] のような質疑応答システムに適している．

*3 検索キーワードの入力時に検索システムから検索キーワードの候補を提示する機能をサジェスト，Web ブラウザによって過去に入力した文字を提示する機能をオートコンプリートと区別する．

■参考文献

1) E.M. Voorhees and D. Harman：“Overview of the Eighth Text REtrieval Conference (TREC-8),”Proc. of the Eighth Text REtrieval Conference (TREC-8), pp.1-24, 2000.
2) N. Kando：“Overview of the third NTCIR workshop,”Working notes of the Third NTCIR Workshop Meeting, pp.1-16, 2002.
3) 北　研二, 津田和彦, 獅々堀正幹：“情報検索アルゴリズム,”共立出版, pp.3-22, 2002.
4) A. Robertson and P. Willett：“Application of n-grams in Textual Information Systems,”Journal of Documentation, vol.54, no.1, pp.48-69, 1998.
5) W. Frakes and R. Baeza-Yates：“Information Retrieval: Data Structures and Algorithms,”Prentice-Hall, pp.28-43, 1992.
6) K. Matsuda and T. Fukushima：“Task-Oriented World Wide Web Retrieval by Document Type Classification,”Proc. of the Eighth International Conference on Information and Knowledge Management, pp.109-113, 1999.
7) H. Kawai, S. Akamine, K. Kida, K. Matsuda, and T. Fukushima：“Development and Evaluation of the WithAir Mobile Search Engine,”in Poster Proc. of the International World Wide Web Conference, 2002.
8) D. Dulitz, A.A. Verstak, S. Ghemawat, and J.A. Dean：“Duplicate document detection in a web crawler system,”U.S. Patent 7627613, 2009.
9) S. Lin and J. Ho：“Discovering Informative Content Blocks from Web Documents,”in Proc. of the Eighth International Conference on Knowledge Discovery and Data Mining, pp.588-593, 2002.
10) J. Zobel, A. Moffat, and K. Ramamohanarao：“Inverted files versus signature files for text indexing,”ACM Trans. on Database Systems, vol.23, no.4, pp.453-490, 1998.
11) P. Indyk and R. Motwani：“Approximate nearest neighbors: Towards removing the curse of dimensionality,”in Proc. 30th ACM Symposium on Theory of Computing, pp.604-613, 1998.
12) G. Salton, E.A. Fox, and H. Wu：“Extended Boolean Information Retrieval,”Comm. of the ACM, vol.26, no.11, pp.1022-1036, 1983.
13) B.J. Janse, A. Spink, and T. Saracevic：“Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web,”Information Processing & Management, vol.36, no.2, pp.207-227, 2000.
14) “インターネット検索エンジンの現状と市場規模等に関する調査研究,”総務省情報通信政策研究所調査研究報告書, 2009.
15) L. Page, S. Brin, R. Motwani, and T. Winograd：“The PageRank Citation Ranking: Bringing Order to the Web,”Technical Report 1999-66, Stanford InfoLab, 1999.
16) J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tonkins：“The Web as a Graph: Measurements, models and methods,”Proc. of the International Conference on Combinatorics and Computing, LNCS.1627, pp.1-18, 1999.
17) G. Navarro：“A Guided Tour to Approximate String Matching,”ACM Computing Surveys, vol.33, no.1, pp.31-88, 2001.
18) 住田一男, 樽井伸司：“多言語横断検索技術について,”特技懇, No.252, pp.71-82, 2009.
19) D. Sullivan：“How Google Instant's Autocomplete Suggestins Work,”http://searchengineland.com/how-google-instant-autocomplete-suggestions-work-62592, 2011.
20) G. Salton and C. Buckley：“Improving Retrieval Performance by Relevance Feedback,”Journal of the American Society for Information Science, vol.41, no.4, pp.288-297, 1990.
21) 細野公男, 緑川信之, 岸田和明（共訳）：“情報検索の認知的転回-情報捜索と情報検索の統合,”丸善, 2008.
22) D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A.A. Kalyanpur, A. Lally, J.W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty：“Building Watson: An Overview of the DeepQA Project,”AI Magazine, vol.31, no.3, pp.59-79, 2010.

■S3 群－3 編－8 章

8-3 ナレッジマネジメント

（執筆者：山口高平）［2018 年 11 月 受領］

1990 年代半ばに，野中郁次郎により，共同化（Socialization），表出化（Externalization），連結化（Combination），内面化（Internalization）という 4 つのプロセスをスパイラル的に繰り返して，暗黙知と形式知が相互に変換されることにより，組織知が創造されていくという SECI（セキ）モデルが提唱された後，1990 年代後半には，トーマス・ダベンポートによって，情報技術を重視したより実践的なナレッジマネジメント（Knowledge Management，以下 KM と略記）の方法論が彼の著書「ワーキング・ナレッジ」で説かれ，経営手法としての KM が広く認知されるようになった [1].

KM を技術的視点から捉えれば，知識の収集，蓄積，更新，分配，共有，創出などを含む一連の知識操作プロセスであり，そのプロセスを支援する情報技術としては，1990 年代後半は，検索エンジン，データベース管理システム，電子メールのような広く普及したツール，あるいは，ワークフロー，ERP（Enterprise Resource Planning），SCM（Supply Chain Management）のような企業情報支援システムが代表的であった.

しかしながら，暗黙知を含む知識を人々の間で共有し，人々が新しい知識を創出していくには，AI 技術が必須の時代になりつつあり，特に，知識表現とデータマイニングとの関連性が深い．以下，各事例について述べる.

8-3-1 知識表現と KM

現場で有効な業務ノウハウは，マニュアルでは記載されておらず，属人化されているケースが多く，8-1 節で述べた ES 開発過程と同様に，専門家にインタビューして専門知識を表現し，KM を開発することになる．文献 2）では，高速道路設備（**図 3･1**）の保守業務知識の獲得と共有を目的として，作業手順を言及したワークフロー，状況判断ルールベース，設備や保守業務に関連する専門用語を体系化したオントロジー，及び設備や業務の理解を助けるための写真・動画に基づく保守業務 KM システムを開発した，開発後，30～40 名の業務担当者が集合し，システム評価会が開催された．その結果，点検現場の判断を反映した状況判断ルールベース，点検業務撮影動画，設備写真など，具体的な知識データは高く評価されたが，具体的作業手順を統合し汎化した，抽象度の高いワークフローは，あまり評価されなかった．設備点検では，知識が設備と一体化しており，その固有知識･データの提示が，業務担当者にとって有用だったと言える．ただ，上長だけが，抽象度の高いワークフローを評価した．「私は，長年，様々な設備点検を経験してきたので，具体的な知識や情報にはあまり関心はないが，抽象度の高いワークフローは，今までの体験を集約化したようなものであり，高速道路設備点

図 3・1　高速道路設備

検業務全体の見直しを考えるうえで有用である」と評価された．以上のように，ユーザに依存して，KM システムで提供する知識を変える必要があると言える．

8-3-2 データマイニングと KM

2000 年前後はデータマイニング，近年はデータサイエンスという名のもとで，データから有用なパターン（知識）を見出すという，知識創出・知識発見的な KM に大きな関心が寄せられている．データマイニング過程は，データ前処理（データ洗浄，属性選択，属性追加など），マイニング（統計処理，機械学習，深層学習アルゴリズムの実行によるモデル構築），データ後処理（マイニング結果の評価）という 3 過程に分かれ，その開発コストは，6 : 1 : 3 もしくは 7 : 1 : 2 と指摘され，マイニングではなく，データ前処理とデータ後処理が重要になっている．これは，どういうことであろうか？　データマイニングでは，現在，マイニングを実装するためのソフトウェアライブラリが充実し，様々なマイニング手法の自動実行が可能となり，開発コストは低くなっている．その一方，データ前処理では，属性選択・属性生成は，対象問題のモデリング過程と同様であり，一部自動実行も可能であるが，自動実行は効果が小さいケースが多く，人手によりデータの意味を考えながらデータの組合せを考える実態があり，コストを要している．また，データ後処理では，人手によりマイニング結果をビジネスゴールの観点から意味を考え多様に評価するため，やはりコストを要している．

10 数年前，サッカーデータマイニングを経験した．**図 3･2** のような特徴量を利用してサッカーボールの移動を時系列データで表現し，プレイヤーの行為（キック，ファウルなど）を追加して，ゴールになりやすいサッカーボール移動パターンをマイニングした．コーチはその結果を高く評価したが，プレイヤーには，「こんな移動パターン，自分達のプレースタイルじゃないよ」と一蹴された．「でも，客観的には，これが勝ちパターンですが」と食い下がったが，相手にされなかった．

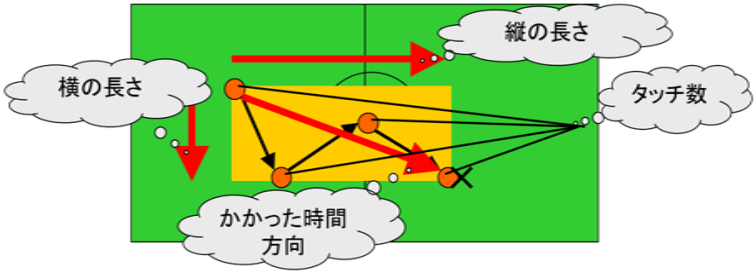

図 3・2　サッカーデータマイニング

現場では，客観（マイニング結果）は主観（担当者の意見）が対立する場合があり，マイニング結果を慣習や業務に摺合せないと，マイニング結果が採用されないケースがある．

しかしながら，ネットベンチャーにおけるビッグデータマイニングは，様相が変わってきたことも事実である．例えば，オンラインゲームなどを提供するネットベンチャーでは，データサイエンティストがゲームクリエイターと対等に意見交換するという．場合によっては，ゲームクリエイターを超えるときさえあるという．客観が主観に勝るのである．例えば，集客数

（KPI : Key Performance Indicator）が落ちたとき，過去のマイニングパターンに基づいてゲームシナリオを変更し，その結果集客数が増えれば，データサイエンティストの意見が採用されるのである．

ただ，リアル企業の状況はさほど変わらない．オフラインでは，現体制を容易に変更はできないので，現場担当者が納得しなければ，データサイエンティストの意見は採用されないのである．決定権を持っているのは，あくまで現場担当者である．データサイエンティストがいくら高精度で客観的な分析結果を提示しても，現場担当者の主観が優先される事態は変化していない．マイニングは，潜在する相関関係を外在化するのであり，因果関係を提示しているわけではない．人は，幼児期に，「お月様は，今は大きくてオレンジ色だけど，どうして，白く小さく変わってしまうの？」など，大人を質問攻めにして，大人が答えをちょっとはぐらかすと，因果関係から鋭く攻撃されることがあるが，これは，幼児期から，人は因果に関する説明を求めるという生来の性質にあるからとされる．このように，相関関係では人は腑に落ちないのである．相関関係から因果関係を再解釈する必要があり，これは，データサイエンスの課題ではなく，知識ベースの新しいチャレンジと言える．

■参考文献

1) T.H. Davenport and L. Prusak：“Working Knowledge: How Organizations Manage What They Know,” Harvard Business School Press, 1998.（梅本勝博（訳）:“ワーキング・ナレッジ :「知」を生かす経営,” 生産性出版, 2000.
2) R. Nambu, T. Morita, and T. Yamaguchi:“Integrating Smart Glasses with Question-Answering Module in Assistant Work Environment,” The Review of Socionetwork Strategies, vol.11, no.1, pp.1-16, 2017.

■S3 群－3 編－8 章

8-4 社会インフラオペレーション

（執筆者：福島俊一）［2016 年 3 月 受領］

人々の生活や社会経済活動を支える社会インフラには，ダム・砂防などの国土保全・水利系，水道・エネルギー施設などのライフライン系，道路・鉄道・空港などの交通運輸系が含まれる [1]．これらにひとたび事故が発生すれば社会は大きなダメージを被るが，今後，社会インフラの老朽化と労働人口の減少が相俟って，そのリスクは高まる [1),2)]．センサや IoT（Internet of Things）を活用した状態監視は行われているものの [3)]，安全で品質・信頼性の高い社会インフラを維持するオペレーションには，状態監視にとどまらず，その結果の判断とアクションにまで踏み込んだ最適化・効率化が必要である．社会インフラの状態監視データは膨大かつ多様で，人手による分析では限界があり，最適オペレーションの実現に向けて人工知能技術が活用され始めている．本節では，そのような人工知能技術の活用事例を 2 つ紹介する．

これらの事例では，人工知能技術として機械学習が使われているが，解釈性の高い機械学習技術を用いていることが特長である．深層学習（Deep Learning）などのニューラルネット系の機械学習技術はブラックボックス型，つまり，判断理由を説明することができないものである．しかし，社会インフラのオペレーションにおいて，理由の説明できないアクションを実行することは信頼性・安全性の面で難がある．それ故，本節で紹介する事例では，理由を説明できる（解釈性の高い）機械学習技術が採用されている．

8-4-1 異種混合学習技術・予測型意思決定最適化技術による上水道オペレーションの事例

上水道のオペレーションにおける課題は，利用者の需要を満たすことを保証しつつ，浄水・配水などで使う電力をできる限り抑えることである．しかも，水道管の劣化による漏水があれば，その箇所を検知するとともに，配水時の水圧を抑えて，劣化した水道管の延命を図ることが求められる．この課題を解決するためには，振動センサなどを用いた漏水箇所の特定だけでなく，いつどこでどれくらいの水量が使われるかのきめ細かな需要予測と，そのような条件下での配水計画の最適化が必要になる．

このきめ細かな需要予測を可能にしたのが異種混合学習技術である [4)]．機械学習技術を用いた水の需要予測では，過去の水使用量の実績とそのときの天候やカレンダー情報などから規則性を発見し，その規則性に基づいて将来（例えば数時間後や 1 日後）の水使用量を予測する．予測精度の高さに加えて，従来の機械学習技術で課題だった解釈性と大規模展開容易性という 2 点を実現したことが，異種混合学習技術の特長である．

その 1 点目，従来技術では，高い予測精度が得られたとしてもブラックボックス型または非線形の予測式を用いたものであったため，予測理由が説明できなかった（解釈性が低い）．異種混合学習技術では，平日/休日，晴/曇/雨のようにいくつかの条件で場合分けし，休日で晴の時間帯は予測式 A，平日で雨の時間帯は予測式 B というように，その場合ごとに予測式（しかも線形などの分かりやすい関係式）を切り替えることで，高い精度と解釈性の両立を実現した．更に，どのような場合分けがよいかと，それぞれの場合にどのような予測式がフィットするかを，独自の情報量基準に基づいて自動決定する．その結果，特長の 2 点目として挙げた大規模展開容易性が得られる．つまり，家庭ごと・オフィスごとの大量できめ細かな需要予測を行う

際，従来の機械学習技術ではいちいち学習パラメータを人手で設定し直すような手間がかかったが，異種混合学習技術ではそれが不要である．

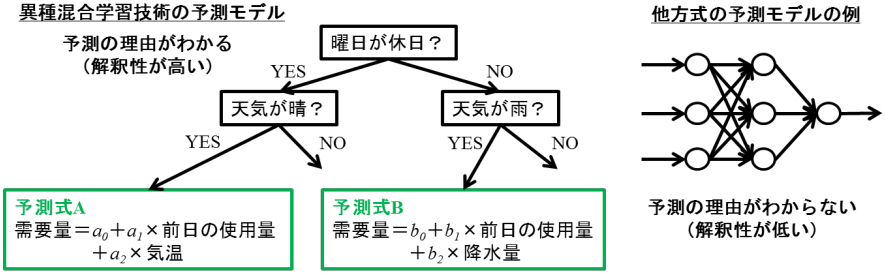

図 4・1　異種混合学習技術の予測モデル

続いて，配水計画の最適化までを可能にしたのが予測型意思決定最適化技術である [5]．異種混合学習技術で得られた各家庭・各オフィスの水需要の予測結果を満たすようにポンプやバルブを制御する，かつ，その際に水道管の漏水箇所・劣化箇所にかける水圧は低く抑える，更に，ポンプやバルブの制御に必要な電力使用量は少ない方がよい，というような条件で最適な配水計画を求めたい．しかし，一つのポンプやバルブの動きが上記の複数要素に影響を与えることや，予測自体が誤差を持つものであることから，条件に最適な制御手順を計画することは容易ではない．この難問に対して，予測型意思決定最適化技術によれば，影響し合う関係や予測誤差まで考慮したうえで，条件にマッチし，リスクも少ない最適計画を高速に探索することが可能になった．

これらの技術を上水道オペレーションに適用した場合の効果として，浄水・配水にかかる電力使用量を約 20％削減できるとの試算が報告されている [5]．また，ここでは水道インフラへの適用事例を紹介したが，同様の予測問題・計画最適化問題は，電力インフラや交通インフラを含む様々な社会インフラのオペレーションで発生するものであり，幅広い技術展開・社会貢献が期待できる．

8-4-2 インバリアント分析技術による原子力発電所の異常予兆検知の事例

原子力発電所などの大規模エネルギー施設のオペレーションにおける重要課題は，異常予兆の早期検知による事故の未然防止である．原子力発電所などで異常が発生すれば，甚大な人的・経済的・社会的損失が引き起こされかねない．温度・圧力・振動・流量などの様々なセンサを配置して状態を監視することは当然行われているが，異常をより早期に予兆段階で検知できれば事故を回避できる可能性が高まる．なお，エネルギー施設の大規模複雑性や，センシング結果を解釈して異常の予兆かどうかを判断できるような高度専門人材の不足などが，この課題の背景にある．

インバリアント分析技術は，専門知識を用いずに，人間よりも早期の異常予兆の検知を可能にした [6]．この技術では，2 つのセンサ間で観測値（時系列データ）に不変関係（インバリアント）が成立するかを，全センサ総当たりで調べる．インバリアントとは，一方のセンサの観測

値が高くなると，もう一方のセンサの観測値も高くなるというような連動した動きをする関係のことである．正常時のセンサ観測値からあらかじめインバリアントを学習しておき，その後の観測値でインバリアントが崩れたとき，異常の予兆だと判断する．あるセンサの振る舞いが異常になると，通常，そこを起点としたインバリアントの崩れを周辺に引き起こすので，異常が起き始めた箇所の特定も可能で，解釈性も高い．

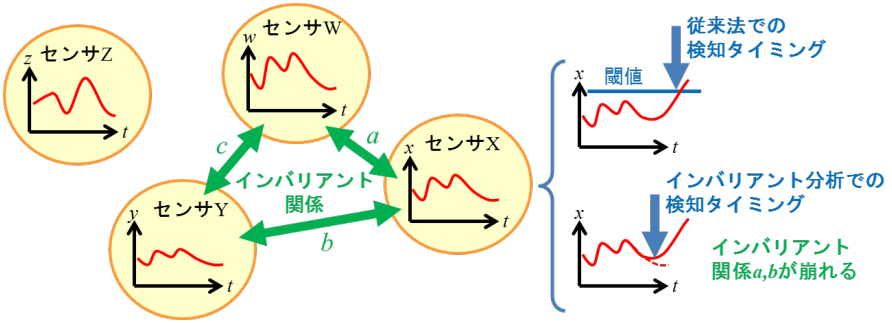

図 4・2　インバリアント分析技術

原子力発電所での実証実験において，センサひとつひとつに閾値を設定する従来法よりも早期の異常予兆検知が可能だったことが報告されている [6]．この理由は，インバリアント分析ではインバリアントが崩れた時点ですぐに予兆を検知できるのに対して，従来法では異常が進行して閾値を超えるまで検知できないためである．また，従来法では各センサの閾値を設定するために専門知識を必要とするのに対して，インバリアント分析ではそれが不要で，設定の手間がかからないというのも特長である．その結果，ほかの社会インフラの異常や劣化の予兆検知への応用も容易で，原子力発電所以外の大規模エネルギー施設や大規模構造物などへの展開も期待できる．

■参考文献

1)　橋本鋼太郎，菊川　滋，二羽淳一郎（編）：“社会インフラ メンテナンス学，”土木学会，丸善出版, 2015.
2)　神尾文彦，稲垣博信，北崎朋希，(野村総合研究所)：“社会インフラ 次なる転換，”東洋経済新報社, 2011.
3)　浅野祐一，木村　駿：“2025 年の巨大市場，インフラ老朽化が全産業のチャンスに変わる，”日経 BP 社, 2014.
4)　藤巻遼平，本橋洋介：“分析プロセス自動化・標準化への挑戦―実践に基づく考察，”情報処理学会デジタルプラクティス論文誌, vol.3, no.6, pp.198-206, 2015.
5)　“NEC，ビッグデータ分析・予測に基づき判断や計画を最適化する人工知能（AI）「予測型意思決定最適化技術」を開発，”（プレスリリース）, http://jpn.nec.com/press/201511/20151102_03.html（2015.11.2）
6)　嚢田昌尚，落合勝博，朝倉敬喜，林　司：“インバリアント分析技術の大規模物理システムへの適用―原子力発電所の監視への適用を例に―，”情報処理学会デジタルプラクティス論文誌, vol.3, no.6, pp.207-214, 2015.

■S3 群－3 編－8 章

8-5 知的教育システム

（執筆者：柏原昭博）［2009 年 3 月 受領］

8-5-1 知的教育システムとは

知的教育システムとは，人工知能技術やメディア技術を基盤に，教師役となるコンピュータと人間の学習者との双方向的かつ適応的なインタラクションを実現するシステムのことである [1]．その目的は，学習者への効率的・効果的な知識の伝達や，学習者による知識の構築・発見を助長することにある．特に，個別的な学習者への対応に力点が置かれ，如何に個々の学習者にとって適切と考えられる情報（演習問題，ヒント，解説など）を適切なタイミングで提示するか，如何に学習者にとって適した学習場や学習ツールを提供するかが，システム開発における主要な研究課題となっている．

知的教育システムは，一般に (1) 教授知識，(2) 教材知識，(3) 学習者モデル，(4) ユーザインタフェース，の 4 つのモジュールから構成される．学習者に対する個別的適応能力を有するためには，学習者の問題解決や理解の状態をシステムが把握する必要がある．そのためには，システム自体が問題を解いたり，理解する能力を備え，そのうえで学習者による問題解決・理解を評価・モデル化する能力を持つことが前提となる．

教材知識モジュールでは，教育対象となる知識を表現するとともに，知識を用いてシステムが問題解決・理解する能力を具備する．学習者モデルモジュールは，ユーザインタフェース上における学習者の入力から，教材知識モジュールの問題解決能力を用いて学習者の問題解決や理解状態をモデル化する．教授知識モジュールは，学習者モデルの情報に基づき，個々の学習者に適したヒントや説明，演習問題を提示することで学習者とのインタラクションを制御する．ユーザインタフェースモジュールは，教授知識モジュールによって決定された情報の提示や，学習場・学習ツールの提供を行うことで学習者とのインタラクションの円滑化を図る．

8-5-2 歴史的変遷

知的教育システムの研究開発は，それが依拠する学習理論や学習観の歴史的変遷の影響を受けて進展してきているといえる．そもそも行動主義的学習観に依拠する CAI（Computer-Assisted Instruction）における様々な問題や限界に対して，認知的学習観に立脚したシステムとして ITS（Intelligent Tutoring Systems）[1] が 1980 年代から 1990 年代にかけて盛んに研究・開発されるようになった．ITS では，個々の学習者に適した教育的指導をコンピュータ化するために，学習者の認知プロセスに焦点を当て，特に学習者の誤り，及び誤りを起こす過程をモデル化し，それに基づいて誤りを修正する教育を実施することに力点が置かれた．個別指導の本質が学習者モデリングにあるという認識が分野全体に定着したのも，主に ITS の研究開発を通してである．

一方，誤りの修正を中心とした知識伝達に力点が置かれることが多い ITS に対して，1990 年代頃から学習者自身による知識の構築こそ知的教育システムにおいて支援すべきであるとの考えが強調され，それと同時に構成主義的学習観，特に社会構成主義的学習観をシステム化する動きが活発になりはじめた．このような学習観に依拠したシステムとして，ILE（Interactive Learning Environments または Intelligent Learning Environments）[2] や CSCL（Computer-Supported

Collaborative Learning）の研究開発が進められた．

ILE では，ユーザインタフェース上において，システムの提供する物理的な対象物を学習者自身が直接操作したり，教材が与える情報空間を探求するための学習環境やツールを提供することで，学習者自ら知識を構築したり，知識を発見することをガイドする．対象物に対する直接操作と対象物の状態変化の可視化を通して，学習者による知識の構築や発見を支援する環境は，特にマイクロワールドと呼ばれる．マイクロワールドは，主に学習者が立てた仮説を検証するための実験環境を提供し，学習者による試行錯誤を促進することで発見的学習を支援するものである．このように，ILE では学習者による学習環境との積極的なインタラクションを通して，学習者自身による知識構築・知識発見を助長する点が ITS とは大きく異なる特徴となっている．

CSCL では，複数の学習者の間で問題や課題を共有し，コンピュータネットワークを介した同期的あるいは非同期的なインタラクションを通じて協調的に知識構築を行うプロセスを支援する．学習者間のインタラクションでは，ある学習者が教える立場でほかの学習者に対して知識伝達を行ったり，対等な立場で問題や知識について議論するといったことが行われる．いずれの場合でも，協調的な知識構築を促進するために，相互作用を交わす学習者を選定したり，各学習者の役割を割り当てるなどインタラクションの場を制御することがシステムの主なタスクとなる．また，協調的な知識構築に加えて，学習プロセスや学習結果を共有することで，学習者のメタ認知を育成することを目的としたシステムもある．

以上のように知的教育システムは変遷してきたが，現在も ITS，ILE，CSCL は様々なドメインを対象として研究開発が進められている．例えば，言語学習では，学習者による言語獲得を支援するために言語に関する基礎知識の伝達を ITS として実現したり，コミュニケーション能力の向上を ILE や CSCL として支援している．また，Web のような大規模なハイパー空間における学習を対象として，Adaptive educational hypermedia systems と呼ばれる ILE の枠組みを用いた適応的な支援の試みがなされている．更に，モバイル・ユビキタス機器などの新しいデバイスを用いて協調的学習の場を拡張し，学習者間のインタラクションを支援する CSCL（mLearning 環境あるいは uLearning 環境と呼ぶこともある）などの研究開発も盛んに行われるようになっている．

8-5-3 限界と課題

知的教育システムの質は，学習者の知識状態や理解状態をモデル化する能力に大きく依存する．しかしながら，これは必ずしも学習者モデルの単なる正確さが重要であることを意味するものではない．学習者モデルでは，効果的な教育を展開するために必要かつ十分な学習者に関する情報が得られるかどうかが本質的に重要である．通常，知的教育システムをデザインする上で，学習者による問題解決や学習のモデルを想定し，そのうえで学習者が直面する困難さを見出し，教育目的・方略を考慮しつつ学習者モデルとして収集すべき情報やその収集手法を決めることになる．しかしながら，制約のあるユーザインタフェースを介して得られる学習者の入力から収集すべき情報が得られない，あるいは推定できないこともあり，必ずしも的確な学習者モデルを構築することは容易ではない．このことは，知的教育システムの質向上を行ううえで非常に重要な問題である．

このような問題に対する一つのアプローチとして，Open Learner Modeling という考え方が注

目されている[3]．これは，システムが生成する学習者モデルを学習者にオープンにして，学習者が学習者モデルに修正を加えたり，新たな情報を追加できるようにするものである．これは，学習者モデルの的確さをインタラクティブに向上させる試みであると同時に，学習者が自分自身の問題解決や理解の状態をリフレクションすることを可能とするもので，高い学習効果を期待することができる有望な手法と言える．

■参考文献

1) E. Wenger, et al.：“Artificial Intelligence and Tutoring Systems,” Morgan Kaufmann, 1987.
2) T. O'Shea, C. O'Malley, and E. Scanlon：“Magnets, Martians and Microworlds: learning with and learning by OOPS,” International Journal of Artificial Intelligence in Education, vol.1, no.3, pp.11-25, 1990.
3) S. Bull and J. Kay : “Student Models that Invite the Learner,” in: “The SMILI Open Learner Modelling Framework,” International Journal of Artificial Intelligence in Education, vol.17, no.2, pp.89-120, 2007.

■S3 群－3 編－8 章

8-6 バイオインフォマティクス

（執筆者：阿久津達也）［2008 年 10 月 受領］

バイオインフォマティクス（Bioinformatics）は生命情報科学などと訳され，文字どおり，生物学と情報科学の学際領域の学問分野である．バイオインフォマティクスでは，ヒトの DNA 配列（DNA Sequence）をはじめとする各種分子生物情報データの情報解析手法の開発，及び，それらの手法を用いた新たな生物学的知識の発見が主な目的とされている．また，情報解析手法の開発と相俟って，これまでに蓄積された DNA 配列，アミノ酸配列，タンパク質立体構造，遺伝子発現データ，タンパク質間相互作用データ，代謝ネットワークデータなど多様，かつ，大量のデータを格納するために，様々な種類のデータベースが開発されてきた．これらのデータベースの多くは WWW を通じて（アカデミック利用には無償で）公開されている．また，単にデータを格納するのみならず，様々な検索，予測などの機能を備え，かつ，ほかのデータへのハイパーリンクなども備えているものの多く，一種の知識ベースとなっている．

8-6-1 分子生物学データベース

これまでに分子生物学に関する様々なデータベースが構築されてきたが，代表的と思われるデータベースを**表 6･1** に示す．

表 6・1　主な分子生物学データベース

種　類	名　称	アドレス（URL）
DNA 配列	GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
DNA 配列	EMBL	http://www.ebi.ac.uk/embl/
DNA 配列	DDBJ	http://www.ddbj.nig.ac.jp/
タンパク質配列	UniProt	http://www.ebi.ac.uk/uniprot/
タンパク質立体構造	PDB	http://www.rcsb.org/pdb/home/home.do
化合物	PubChem	http://pubchem.ncbi.nlm.nih.gov/
化合物	LIGAND	http://www.genome.jp/ligand/
文献	MEDLINE/PubMed	http://www.ncbi.nlm.nih.gov/sites/entrez
タンパク質配列分類	Pfam	http://pfam.janelia.org/
タンパク質配列分類	COG	http://www.ncbi.nlm.nih.gov/COG/
立体構造分類	SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
代謝ネットワーク	KEGG	http://www.genome.jp/kegg/
タンパク質相互作用	DIP	http:/dip.doe-mbi.ucla.edu/
遺伝子オントロジー	GO	http://www.geneontology.org/

最も基本的なデータである DNA 配列を格納するデータベースは，米国の GenBank，欧州の EMBL データベース，日本の国立遺伝学研究所の DDBJ の 3 か所に集約されている．論文発表の際には DNA 配列をいずれかのデータベースに登録することになっており，かつ，これらは密接な協力関係にあり，互いにデータを交換し合っている．

DNA 配列は A, C, G, T の 4 種類の文字からなる文字列で表現されるが，タンパク質は 20 種類のアミノ酸からなる文字列で表現される．タンパク質のアミノ配列を格納したデータベースはいくつかあるが，最も広く利用されているのが UniProt データベースである．また，タンパク質の多くは生体内で固有の立体構造（3 次元形状）をとり，この構造は機能に密接に関係している．そこで配列データ同様，タンパク質立体構造データを格納することも重要となるが，それらは PDB（Protein Data Bank）データベースに格納されている．化合物に関するデータは長年にわたって商用のデータベースに蓄積されてきたが，近年では，PubChem や LIGAND などの公的なデータベースにも格納されるようになってきている．

上記のデータベースは実験により得られたデータを格納するものであるが，それらのデータや関連データを解析して得られたデータを格納するデータベースも数多く構築されている．生物学関連の論文情報を集めた文献データベースとして MEDLINE/PubMed がある．タンパク質配列を分類したデータベースも数多くあり，隠れマルコフモデルを用いて分類した Pfam や進化的関係から分類した COG などが有名である．タンパク質立体構造を 3 次元形状の類似性などに基づき分類したデータベースとしては，SCOP などがある．近年では，個々の物質ではなく，ネットワークや相互作用情報を格納するデータベースも数多く構築されており，代謝ネットワークなどのデータを集めた KEGG, タンパク質相互作用データを集めた DIP などが有名である．更に，遺伝子などのデータを対象にオントロジー（Ontology）を定義し，それらに基づき配列の分類を行う GO データベースなどもある．

8-6-2 検索・予測技術

分子生物学データベースの多くは単にデータを格納するだけでなく，様々な検索機能を備えている．配列データを格納するデータベースの多くに類似配列の検索機能が提供されており，生物学者などにより日常的に利用されている．類似配列検索は基本的には動的計画法（Dynamic Programming）に基づく配列アラインメント（Sequence Alignment）アルゴリズムが用いられるが，膨大なデータに対し高速な検索を行うために様々な工夫がなされている．また，データベースに付随して，もしくは，それとは独立に，様々な解析を行う Web サーバも数多く開発され，公開されている．複数の配列を同時に比較するマルチプルアラインメント計算，配列からの進化系統樹の推定，タンパク質配列や RNA 配列からの立体構造の予測，タンパク質配列からの機能の予測，遺伝子発現データからの制御ネットワーク推定など様々な予測サービスを行うサーバが開発されており，それらには，隠れマルコフモデル（Hidden Markov Model）やベイジアンネットワーク（Bayesian Network）などの確率モデル，動的計画法やシミュレーテッドアニーリングなどの最適化技術，サポートベクターマシン（Support Vector Machine）などの機械学習技術など，人工知能における様々な技術や方法論が応用されている．

紙面の関係で簡単にしか説明できなかったが，バイオインフォマティクスの様々な事項については文献 1）を，日本で開発されているデータベースやソフトウェアについては文献 2）を参照されたい．また，Nucleic Acids Research という学術誌には毎年，データベース特集号，Web ツール特集号が出るので，バイオインフォマティクスに関するデータベース及びソフトウェアの最新の情報や動向を得るのに有用である．

■参考文献

1) 日本バイオインフォマティクス学会(編)：“バイオインフォマティクス事典,” 共立出版, 2006.
2) 森下真一, 阿久津達也(編)：“バイオデータベースとソフトウェア最前線,” 実験医学（増刊号）, 2008.