解　説

大規模データベースにおける知識獲得†

西尾　章治郎††

1. はじめに

世はまさに情報化社会の時代である．コンピュータ技術の著しい発展とハードウェアの低価格化にともない，最近では，日々刻々と多種多様な膨大な量の「生」データが十分に解析されないまま，データベースに格納されている．ある統計では，20カ月ごとに世界の情報量は2倍に膨れ上がり，さらに今後も，医療，商業，経済，科学技術，工業などさまざまな分野のデータベースにおいて，データの質・量の両面から着実な拡張が見込まれる．特に，米国航空宇宙局（NASA）などでは，蓄積されているデータの量は，すでに解析可能な量を超過していると言われている．実際，宇宙関係をはじめ，人ゲノム関係のプロジェクト，より日常生活に近いものとしては国勢調査データの格納などのために用いられているデータベースでは，全データの数が数百万から数十億（つまり，テラバイト単位）も格納された**大規模データベース** (Very Large Database ; VLDB) が構築されている．

データは，「知識」の根源である．しかし，データを大量に保有していることと有用な知識を多くもっていることとはまったく別のことである．大量な共用データを有効利用するためには，格納されたデータをよく理解し，有用な知識を迅速かつ的確に発見する必要がある．ただし，ここでの知識とは，より具体的にはデータの領域間に内在する**規則性** (regularity) とか，異なる属性の値に関して成立する「IF THEN ルール」などを意味する．これら獲得された知識を，たとえば，**事実データ** (fact data) のみより構成される従来のデータベースに付加することにより，

1. データベースへの問合せ処理能力の強化
2. 推論エンジンとしての機能の装備

をすることが可能になり，さらには，ルールの数が数千もあるような高度で共用の**大規模知識ベース** (Very Large Knowledge Bases ; VLKB) の構築が見込まれる．

このように大規模データベースのデータ量に関する現実とその有効利用を目指して，**データベースにおける知識獲得** (knowledge discovery in databases) の研究が最近注目を集めている．たとえば，**機械学習** (machine learning) の分野の研究の牽引者である D. Michie は，これまでの機械学習の研究および開発されたツールを今後いかに大規模データの解析に有用化するかが重要な課題であることを強調している[17]．一方，データベースの分野においてもこの分野の研究の重要性が指摘されている．たとえば，米国科学財団 (NSF) で企画された 1990 年代のデータベースの分野の研究テーマを探るワークショップにおいて，データからの**知識発掘** (knowledge mining) の研究は最重要研究テーマの一つにランクされている[27]．最近では，VLDB 国際会議など世界的によく知られたデータベース関連の学会でも，このテーマに関するセッションが設けられるようになった．さらに，人工知能関連の国際会議と連動して，データベースにおける知識獲得の研究のみにテーマを絞ったワークショップがすでに 1989 年と 1991 年に開催された[19),21]．本稿では，データベースにおける知識獲得に関して，従来の人工知能における機械学習の研究などとの関連，最近の研究動向と今後の展望を中心として概説することにする．

ここで，大量データからの知識獲得という言葉には，大きく二つの意味があることに注意しなければならない．最初の意味は，**科学的発見** (scientific discovery) に関連するもので，実験データのように対象世界で観測される大量の（通常は，数

† Knowledge Discovery in Very Large Databases by Shojiro NISHIO (Faculty of Engineering, Osaka University).
†† 大阪大学工学部

値）データから，最終的に導き出したい規則性を想定しながら，注目しているパラメータ値に関する標本データを知識獲得の対象とするものである．たとえば，気体に関する大量の実験データから圧力，体積，温度に関するデータを抽出し，そのデータを対象としてボイル・シャルルの法則を導出するようなことである．その場合，おのおののデータは標本であり，1個のデータそれ自体はあまり重要な意味をもたない．したがって，法則の導出がうまくいかない場合など，実験を再度実行して標本データを採り直すこともありうる．それらのデータを全体的にみた場合の構造特性，データ間の関数関係などを発見し，標本化の統計的性質も考慮しながら対象世界で成立しているデータ間の規則性を推定するのが研究目的であり，人工知能技術の統計推定への支援が大きな課題となる．このような科学的発見のための実システムとしては，P. Langley らによって開発されたBACON システム[14]が知られている．

もう一方は，ビジネスデータのように，おのおののデータが現実を反映した貴重なデータである場合であり，そのような事実データを大量に格納した大規模データベースが与えられているという前提から出発し，そのデータベースに潜んでいる規則性やルールを発見する研究である．このようなデータベースには，それを利用する組織内でのさまざまな目的に対処できるように，非常に多種類の情報が格納されるのが通常である．したがって，データ間に成り立つ規則性やルールも一見そのような多様な情報のなかに埋もれてしまいがちで，最終的に得られる結果に関してある程度目標を定めて知識獲得を行うことは困難である．もし，目標が定まったとしても，科学発見の場合と異なり，それを導出するために対象となるデータを入力し直すことも通常は行われない．また，更新などのデータベースシステムならではの操作も考慮する必要があり，人工知能技術とデータベース理論を融合した応用が課題となる．本稿では，特に断わらない限りは後者の立場からの議論を行う．

2. 人工知能研究の応用技術として

データベースにおける知識獲得の研究を行う場合に，対象とするデータベースが単純ならば，ばらばらなデータを正確に一般化したり，一見したところ無秩序に見えるデータからある規則性を見い出すことができる．ところが，データが大量かつ複雑になるにつれ，このような知識を獲得するためにコンピュータを使って人間の作業を支援させたり，コンピュータに人間の代役を任せたりしなければ，データを簡単には処理できなくなる．つまり，機械学習の分野での研究成果，開発された技術を有効に利用することが重要になってくる[18]．もちろん，人間の能力をそのままコンピュータに移植することは困難ではあるが，コンピュータによる知識獲得のための実用可能なシステムはある程度実現している．たとえば，少数ながらもすでに医療，CAD/CAM，化学，株の売買などの分野では，1980年代の前半からデータベースから有用な知識を獲得する研究は開始されており（文献 8）を参照），実用システムも開発されている（例として，文献 22）参照）．また，大量データ内に存在する規則性を学習する方法もいくつか提案されている（文献 2) 参照）．このような状況下で実用的な知識発見のためのシステムが開発され，応用されていく可能性は十分にあると考えら

表-1　データベース管理と機械学習の対象の相違

データベース管理の立場	機械学習の立場
データベースは動的で，時々刻々更新されている．	データベースは単なるデータの集合で，変化のない静的なものである．
レコード値は，不完全であったり，誤り情報を含む可能性がある．	事例は，通常完全であり，ノイズは含んでいない．
フィールド値は，通常数値である．	事象列は，通常2値である．
データベースは，通常数百万個のレコード値を含む．	データ集合は，通常数百個のインスタンスを含む．
人工知能研究は，もう少し現実を反映すべきだという考え方が多い．	データベースに関する問題はすでに解決済みであり，研究テーマとしては興味がない．

れる．すなわち，データベースから有用な知識の金塊（nuggets）を的確に掘り起こす機械学習技術を研究・開発する絶好の機会を迎えているといえる．

大規模データベースの管理システムが対象とするデータベースと従来の機械学習の研究が対象としてきたデータベースの相違点を，文献 8）も参考にしながら要約すると**表-1** のようになる．この表を参照しながら，機械学習の理論を応用してデータベースにおける知識獲得の研究を推進するうえで，その対象がデータベースシステムに格納された大量かつ多様な生データであることから特に留意しなければならない三つの問題点を述べることにする．

2.1 アートから工学へ

従来の人工知能の研究，特にその応用であるエキスパートシステムの研究・開発においては，人間のもつ知識の構造に迫ることを主要な目標とした．その目的を達成するためには，通常，問題あるいは応用の対象領域を狭く限定し，その領域に関して利用可能な**背景知識**（domain knowledge）を最大限に利用して，さらに深い知識へと掘り下げ，最終的には知識の核心に少しでも接近しようという方法が取られてきた．このような方法論はある程度の成功を収めてきたが，一般性に欠けるという問題点を含んでおり，ある対象領域に関して開発された方法論が，その領域の特殊さゆえに他の領域には応用できない場合がしばしばあった．大量の生データを対象とする場合には，対象領域そのものを限定することが難しく，領域を狭く限定して知識の核に迫る方法論には限界があり，浅くてもよいから，より広い範囲で一般的に成立する知識の獲得法の開発が必要となる．つまり，より工学的な研究の方法論が要求される．

2.2 正の事例のみからなる学習

一般に帰納学習における**例からの学習（概念獲得）**において，あるクラスに属する事例から帰納的にルールを導く際の理論的基礎として，完全性条件と無矛盾性条件がある[16]．このうち無矛盾性条件は，**正事例**（positive set）および**負事例**（negative set）からの観察をもとに有効な規則性を導くことに関連している．ところが，データベースに蓄えられているデータは正事例（事実）のみであり，負事例（反例）が存在しないのが通常である．そのために無矛盾性条件を適用できず，ルールを一般化するうえでの自然な限界がない（つまり，たとえば，“情報工学科の学生は，人間である”というような，あまりにも一般的で有効に使うことのできないルールが導かれてしまう）．このように正事例のみからの知識獲得を行うために，データベース構築時に必要な属性間の種々の制約や属性値の階層構造・包含関係といった制約を付加することによって，**必要以上のルールの一般化**（over-generalization）を行わないような制約を設ける必要がある．

2.3 非単調な知識の発展

大規模知識ベースのメインテナンスに従事している技術者の間では，「比率 90/10 の法則」ということが言われている．つまり，知識ベース内に蓄えられたデータのうち，規則性に反するたった1割のデータ（ノイズ）のために，全体システムの管理に要する時間の9割を要しているということである[6]．データの更新が多い大規模データベースでは，ある時点で満足されていたルールがデータの更新後に満たされなくなる可能性も多分にあり，今後このような問題には特に留意しなければならない．つまり，少数のデータがルールを満たさないために有用なルールが捨てられるのではなく，例外を扱うための知識の**非単調な発展**（non-monotonic knowledge evolution）を考慮しなければならず，たとえば，完全性条件の緩和などの対策が考えられる．また，人工知能を含めた他の分野におけるこのような問題に対処する技術としては，次に列挙する方法が考えられる[6]．

- 例外処理（exception handler）．
- ルールの動的改訂．
- 再定義（overriding）．
- デフォルト（default）論理．
- 異常宣言述語（abnormality predicate）．
- 確率の導入．計算の複雑度と関連した**PAC学習**（probably approximately correctly learning）（たとえば，文献 26）参照）．
- ファジィ論理．

このような方法をベースとした非単調な知識の発展に関する研究は，今後最も興味深く，現実的に解決が急がれている課題である．

2.4 対象メディアの多様化にともなう課題

人間の知識は，時と場合に応じて，文字，図形，音声などの種類の異なる媒体，いわゆるマルチメディア情報として表現される．ところが，これまでコンピュータに知識ベース化され，エキスパートシステムなどにおいて利用するために知識獲得あるいは知識表現の対象とされてきたのは，文字を媒体とするものがほとんどであった．今後，**マルチメディア**を対象とした大規模データベースの構築が今まで以上に推進すると考えられ，それにともなってマルチメディア情報からの知識獲得が非常に重要な課題になってくると考えられる．ただし，マルチメディア情報の知識ベース化においては，マルチメディアを対象とした知識表現の方法論の確立と，それによって構造化されたデータを格納するための「知識ベースの柔構造化」の研究の推進が重要である．特に，データベースの研究の拡張として，マルチメディア情報を無理に加工することなく，かつ，データ間の複雑な相互関連性を保ちながら蓄積可能にする柔軟なデータベース・スキーマの構築が重要であり，その実現のために**オブジェクト指向データベース**[3]が現時点で最も可能性を備えたシステムと言える．さらに，構造的な柔軟さに加えて知識ベースに必要な演繹機能を併せもったデータベースが望まれ，**演繹オブジェクト指向データベースシステム**のさらなる開発が期待される[29),30)].

3. 組および属性指向アルゴリズム

現在，データベースシステムとして最も普及の著しい**関係データベース**を対象として，コンピュータを使ってルールを自動的に学習するために最近提案されたアルゴリズムのうち次の2種は，データベースの構造を密接に反映したものとして注目に値する．

一つは，関係データベースの組において成立する従属性に注目してルールを求める組指向 (tuple-oriented) のアルゴリズムである．たとえば，**表-2**に示されたような大学に属する学生のデータが与えられたと仮定すると，この表を**組指向**（つまり，行をベース）に従属性を調べると「女性であれば，年齢が 22 歳であり，国文学科に属している」というルールが成立していることが分かる．G. Piatetsky-Shapiro は，組指向のアルゴリズムについて，すべてのデータに対して**並行検索** (parallel check) を行い，しかも，各組へのアクセスを1回しか要求しないような方法を提案している[20)]．このアルゴリズムは，従来から関係データベースの分野において，盛んに行われてきた**関数従属性** (functional dependency) の研究との関連性が非常に深い．

表-2　大学に属する学生のデータ

名前	性別	年齢	住所	所　属
尾崎	男	18	京都市	情報工学科
河野	女	22	相生市	国文学科
津田	男	23	城陽市	数理工学科
柴田	女	22	西宮市	国文学科
松山	男	23	西宮市	英文学科
渡部	男	22	宝塚市	考古学科

もう一方は，属性間の概念的な階層構造に注目してルールを求める**属性指向** (attribute-oriented) のアルゴリズムである．J. Han らは，属性指向のアルゴリズムとして，学習作業に関する知識を**概念木** (concept hierarchy) として与え，属性ごとにその概念木を葉から根に辿っていきながら，各属性の値をより一般的な概念をもつ値に置き換え，その結果成立するルールを導出する方法を提案している[4),10)]．たとえば，表-2 の住所の属性に関して，“京都市”，“城陽市” を “京都府” に，“相生市”，“西宮市”，“宝塚市” を “兵庫県” という一般化した値に置き換え，一方，所属の属性に関しては，“情報工学科”，“数理工学科” を “工学部” に，“国文学科”，“英文学科”，“考古学科” を “文学部” という一般化した値に置き換える．このように属性値の一般化された組に対して，選択・射影を行うことによって，「京都府の学生は工学部に，兵庫県の学生は文学部に属している」というルールが得られる．

通常，獲得されたルールには，データベースに蓄えられたすべてのデータが満足する概念の特徴を表す**特性ルール** (characteristic rule) と，あるクラスの概念を他のクラスの概念と区別する陳述である**分類ルール** (classification rule) が考えられる．たとえば，ある大学のテニス部の学生の特徴を表すルールは特性ルールであり，ある大学のテニス部の学生がその部の顧問の先生と異なったどのような特徴をもつかを表すルールは分類ルールであ

る．J. Han らの研究は，関係データベースを対象として，与えられた概念木からこれら2種のルールを導くアルゴリズムを与えているが，これらのアルゴリズムは，**教示学習** (learning from examples) の一般化規則 (generalization rule) である**条件削除** (dropping condition) 規則，**一般化木上昇** (climbing generalization tree) 規則にもとづいており，同時に**概念クラスタリング** (conceptual clustering) などの手法を併用している．

以上の組指向，属性指向のアルゴリズムでは，対象とする関係データベースの属性は最初に与えられたままを用いたが，その属性自体を変更したり，現在与えられているいくつかの属性の値から計算される属性を新たに追加することにより，なんらかの規則性が得られる場合がある☆．たとえば，科学的発見に関連した大量データからの知識獲得では，気体の“圧力 p”，“体積 v”，“温度 T”に関する大量の実験データがある場合に，新たに pv/T を計算した属性を考えると，この値が（実験データなので）ほぼ一定であるという規則が得られる．ただし，一般的にこのように適切な属性を発見することは，データベースからの知識獲得の研究においては非常に難しい問題である．対象領域に関する背景知識，人工知能技術，統計推定技術などを駆使して，有意義な属性をシステマチックに導き出すための研究が行われている（たとえば，文献 31）参照）．

4. 最近の応用システムからの話題

最近開発されているデータベースからの知識獲得システムのうち，注目に値するいくつかの例を簡単に紹介しよう．

まず，G. Piatetsky-Shapiro は，所属する米国 GTE の研究グループで**知識獲得ワークベンチ** (knowledge discovery workbench) を開発している[22]．通常数値データより構成されるビジネスデータベースを対象として，**対象データの可視化** (visualization) と代表的なデータ解析手法である**クラスタ分析** (clustering)，**分類化手法** (classification)，**要約化手法** (summarization) を用いることでデータ間に成立する規則性を発見するプロトタイプシステムを構築している．ワークステーション上で Common Lisp をベースとして作成されており，今後商用関係データベースとのインタフェースを開発し，実システムへ応用することを目標としている．

最近，米国 MCC で D. B. Lenat らが中心となって開発が進んでいる知識ベース CYC[15] は，大規模知識ベースシステムの今後の構築可能性を探る意味でも非常に重要である．この CYC システムを対象として，W. M. Shen は，$P(x,y)\wedge R(y,z)\Rightarrow Q(x,z)$ のタイプの規則性を発見する手法を開発している[25]．たとえば，“x さんは y さんを非常によく知っている”，しかも，“y さんは z 語を話す”．そのとき，常識的に“x さんは z 語を話す”と考えられ，データ間に成立する多くの興味ある規則性が上記のような形式で記述され得る．このような**常識的な知識** (common-sense knowledge) を知識ベース CYC から発見し，それを付加することによって，より強力な知識ベースの開発が促進されると考えられる．

データベースが大規模化するにつれて，通常フィールド値に種々の誤りが生まれる可能性が高くなる（たとえば，正式には“two”と入力しなければならないところを，システムの大規模化にともなう初歩的な注意不足から誤って“2”と入力してしまうなど）．ワシントン州立大学の J. C. Schlimmer は，このようなデータベースの大規模化にともなって発生する諸々の問題を解決するために，データベースから学習理論を用いてフィールド間に成立する規則性を見い出し，誤り検出を可能にするモデルを構築し，さらに実システム CARPER を開発している[24]．

マルチメディア情報を対象とした研究の重要性を先に述べたが，現時点では知識獲得システムが次々と開発されるほど進展しているわけではない．そのような状況のなかで，遺伝子形状を抽出する研究は活発に行われており，これらの研究は，3次元構造物に関するデータからの知識獲得技術の開発と関連すると考えられる．たとえば，Rutgers 大学の D. Cohen らの研究グループは，DNA の水化パターンを推測する研究を行っているが，そこでは，DNA のクリスタル構造をクラスタ解析と人工知能における**決定木** (decision tree) 技術を応用して類別化することを目指している[5]．方法論としては，まず，クラスタ解析をし

☆従来のデータベースの分野における関連研究として，要約データ（たとえば，文献 12) 参照）に関する研究がある．

てパターンの分類を行ったうえで，特徴抽出を決定木で行っている．このような研究はパターン認識に関する研究ともみなせるが，別の見方をすれば，DNA の構造に関する大量のデータから形状物に関する特性を抽出するという意味で，知識獲得技術に関する研究とも考えられる．

最近，ノードとリンクの概念によって実世界の対象システムの表現を行う**ハイパテキスト** (hypertext) の考え方に基づいたソフトウェアシステムの開発が盛んに行われている．ハイパテキストシステムも大規模化するにつれリンクの構造が複雑化し，その管理とリンク上のナビゲーション自体が非常に複雑になり，ハイパテキストの利点が失われてくる．このような問題に対して，原は，大規模ハイパテキスト構造に関して，概念設計レベルにおけるリンクの抽象化方策と，物理構造レベルおよびヒューマンインタフェースレベルにおけるリンクのクラスタリング手法による解決法を提案している[11]．その方法は，大量のリンクデータを格納したデータベースから大域的情報の抽出をはかる知識獲得の研究と考えることができる．

近年，通信ネットワークの統合拡大による物理的な規模の拡大，および，それにともなう管理に関する論理的な複雑度の急速な増大にともないネットワークマネージメントが困難になってきている．今後も，通信ネットワークが高度化するにつれて，情報伝送・処理システムにおける MIB (Management Information Base) をはじめとする多種類の情報システムが増加し，論理的な構成の複雑さがより大きく性能に影響することは明らかである．河野らは，大規模データベースにおける知識獲得の手法を統合化された大規模通信ネットワークの状態に応じたマネージメントを行う一つの手段として導入することを提案している[13]．

5. 処理性能向上のための課題

全データの数が数百万から数十億にも及ぶような大量の共用データベースからルールを発見するには，非常に効率の良いアルゴリズムであっても，すべてのデータに適用するのは困難であろう．特に，それらのすべてのデータを従来の逐次的な方法で検索していたのでは，いくら強力なコンピュータと高速なアルゴリズムを採用しても計算時間が爆発的に増大してしまい，実用的な時間ではなかなか計算が終了し難い．そこで，現時点でこの問題を解決する方法としては次の二つが有力である．

1. 独立に計算が可能な領域に分けて，**並列計算** (parallel computing) を実行する．

2. 全データを検索せず，**サンプリング** (sampling) をして得られた一部のデータからの情報を用いて有用なルールを導く．

並列計算の実行可能性に関しては，対象データを相互依存性のない領域にいかに分割して，知識獲得アルゴリズムを並列に実行するかが問題となる．そのアルゴリズムを実行するうえで，推論能力を要求されるような高度な計算を必要となる可能もあり，ICOT で開発されてきた大規模並列推論マシンなどの重要な応用領域とも考えられる．

一方，データから任意に取り出したサンプルにアルゴリズムを適用してルールを発見する場合には，サンプルから導出されたルールが，構築された大規模知識ベースにおいて適当であるとは限らない．そこで，サンプルから導出されたルールの大規模知識ベースでの正確さを評価しておく必要が生じる．つまり，確率・統計の理論を用いて，全データを検索した場合との比較のもとで，得られたルールの正確さをいかに定量的に保障するかが重要な課題となる．非単調な知識の発展 (**2.3** 参照) との関連で述べた PAC 学習も一つのサンプリング手法といえる．

この分野の研究として，G. Piatetsky-Shapiro は，組指向のアルゴリズム (**3.1** を参照) を用いてサンプルから導出されたルールが，全データベースに適用されたときの純計的な正確さの解析を行っている[20]．また，園生らは，属性指向のアルゴリズム (**3.2** を参照) を用いてサンプルから導出されたルールが，全データベースに含まれる全体の組に適用されたときの統計的な正確さの解析を行っている[28]．さらに，阿久津と高須は，関係データベースにおける関数従属性の PAC 学習可能性について研究を行っている[1]．

また，従来の決定木学習アルゴリズムを処理性能向上のために有効利用する方法として，ID 3[23] などのよく知られている決定木の手法によって得られた出力を，W. J. Frawley が示している *Function Based Induction Algorithm* にみられるようなルールに変換することも重要である[7]．

6. 今後の展望

ここ数年，データベースと知識ベースとの接点あるいは相違点に関する議論が多くなされている．Y. Freundlich は，それら二つの相違を**表 3** のようにまとめている[9]．この表からも明らかなように，データベースにおける知識獲得の研究は，専門家を仮定しない一般の情報収集機構によって大量に蓄積されたデータを対象として，多目的でより高度な情報システムである大規模知識ベースへ変換していくための核技術として非常に重要である．さらに，「知識の泉」ともいえる大量の生データのなかから貴重な財宝を掘り起こし，情報化社会に還元するための非常に大切な機械学習の応用研究課題として，今後ますますの発展が期待される．

具体的にどのような研究テーマが重要かについて，1991年の知識獲得に関するワークショップで「データベースにおける知識獲得に関するヒルベルト問題」と題する興味深いパネルセッションが行われた．大数学者ヒルベルトが，1900年に解決が非常に困難な数学の問題を「今世紀に残された23の問題」と称して提示したことは周知のとおりである．それにならって，データベースにおける知識獲得に関して，今後解決していくべき重要な問題を列挙し，今後の方向を探るための活発な討論が行われた．その重要テーマのなかには，いろいろな知識獲得技術の融合に関する問題，ユーザインタフェースを考慮した対話的システムの構築の必要性，データに内在する不確かさの扱いに関する問題，オブジェクト指向データベースなどの複雑な構造をもったデータからの知識獲得技術の開発などが含まれている．

日本でも今後推進すべき国家的研究プロジェクトのテーマとして，大規模知識ベースシステムの構築がクローズアップされ，その構築技術に関する議論が盛んに行われている．**図-1** に筆者が考える大規模知識ベースのシステム構成の一例を示した．その図では，単に専門家の知識のみでなく，知識ベースの構成要素である大規模データベースから知識を獲得し，その得られた知識をシステムに有効にフィードバックして内容的により豊富な共用知識ベースを構築する重要性を強調した．今後，特に共用データベースから知識を獲得

表-3 データベースと知識データの相違

比較項目	データベース	知識ベース
情報の収集者	事務員	専門家
利用目的	情報検索	多目的
情報の種類	事実	より高度な情報（知識）
理論的な要請	計算の理論	意味論的な解釈

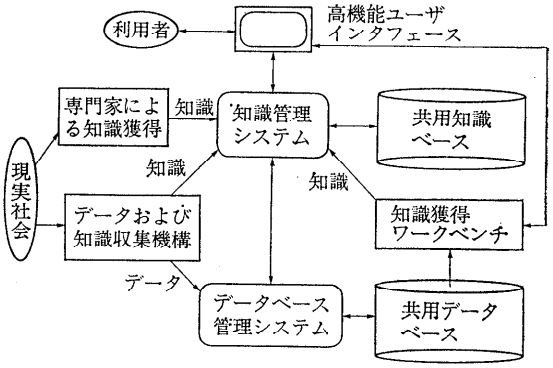


図-1 大規模知識ベースのシステム構成例

する高機能な知識獲得ワークベンチの構築が急務であり，さらにこの構築技術をさらに一般化し，データベースからの**知識獲得工学** (Knowledge Discovery Engineering) へと発展させることが重要であると考える．

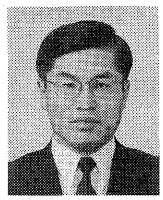
謝辞　末筆ながら，本稿に対して貴重なコメントをいただいた京都大学河野浩之助手に心より感謝の意を表す．また，日頃ご指導いただく京都大学長谷川利治教授および茨木俊秀教授，ならびに大阪大学宮原秀夫教授に深謝の意を表す．なお，本稿の一部は，東京大学大須賀節雄教授を研究代表者とする文部省科学研究費重点領域「知識科学」の補助を得ている．

参考文献

1) 阿久津達也，高須淳宏：関係データベースにおける関数従属性の PAC 学習可能性について，第6回人工知能学会全国大会論文集，pp. 327-330，東京 (June 1992)．
2) 安西祐一郎：認識と学習，岩波講座ソフトウェア科学 16，9章，pp. 294-327，岩波書店，東京 (1989)．
3) Atkinson, M., Bancilhon, F., DeWitt, D., Dittrich, K., Maier, D. and Zdonik, S.: *The Object-Oriented Database System Manifesto*, Kim, W., Nicolas, J.-M. and Nishio, S. (eds.): *Deductive and Object-Oriented Databses*, pp. 223-240, North-Holland, Amsterdam (1990)．
4) Cai, Y., Cercone, N. and Han, J.: *Attribute-Oriented Induction in Relational Databases*, 文献 21)，pp. 213-228 (1991)．

5) Cohen, D., Schneider, B., Berman, H. and Kulikowski, C.: *Learning to Predict DNA Hydration Patters*, 文献 19), pp. 10-122 (July 1991).
6) Esculier, C.: *Non-Monotonic Knowledge Evolution in VLKDBs, Proc. of the 16th International Conference on Very Large Data Bases*, pp. 638-649, Brisbane, Australia (Aug. 1990).
7) Frawley, W. J.: *Using Functions to Encode Domain and Contextual Knowledge in Statistical Induction*, 文献 21), pp. 261-275 (1991).
8) Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J.: *Knowledge Discovery in Databases: An Overview*, 文献 21), pp. 1-27 (1991).
9) Freundlich, Y.: *Knowledge Bases and Databases: Converging Technologies, Diverging Interests, IEEE COMPUTER*, Vol. 23, No. 11, pp. 51-57 (Nov. 1990).
10) Han, J., Cai, Y. and Cercone, N.: *Data-Driven Discovery of Quantitative Rules in Relational Databases*, To appear in *IEEE Transactions on Knowledge and Data Engineering* (1993).
11) 原 良憲：ハイパーテキストデータベースにおける抽象化操作とクラスタリング手法，オブジェクトテクノロジーの高度応用に関する Obase ワークショップ講演論文集，pp. 3-10，神戸 (Jan. 1992).
12) Johnson, R. R: *Modelling Summary Data, Proc. of International Conference on Management of Data* (ACM-SIGMOD 1981), pp. 93-97, Ann Arbor, Michigan (Apr./May 1981).
13) 河野浩之，西尾章治郎，長谷川利治：大規模データベースにおける知識獲得アルゴリズム—通信ネットワークにおける管理知識の獲得—，第6回人工知能学会全国大会論文集，pp. 335-338，東京 (June 1992).
14) Langley, P., Bradshaw, G. L. and Simon, H. A.: *Rediscovering Chemistry with the BACON System*, in Michalski, R. S., et al. (eds.): *Machine Learning: An Artificial Intelligence Approach*, pp. 307-329, Springer-Verlag (1984).
15) Lenat, D. B. and Guha R. V.: *Building Large Knowledge-Based Systems*, Addison-Wesley, Reading, MA (1990).
16) Michalski, R. S.: *A Theory and Methodology of Inductive Learning*, in Michalski, R. S., et al. (eds.): *Machine Learning: An Artificial Intelligence Approach*, pp. 83-134, Springer-Verlag (1984).
17) Michie, D.: *March 15 Interview, AI Week*, Vol. 7, No. 6, pp. 7-12 (1990).
18) 西尾章治郎：大規模データベースからの知識獲得と機械学習，人工知能学会誌，Vol. 7, No. 1, pp. 13-16 (Jan. 1992).
19) Piatetsky-Shapiro, G. (eds.): *Proceedings of 1991 AAAI Workshop on Knowledge Discovery in Databases*, Anaheim, CA (July 1991).
20) Piatetsky-Shapiro, G.: *Discovery, Analysis, and Presentation of Strong Rules*, 文献 21), pp. 229-248 (1991).
21) Piatetsky-Shapiro, G. and Frawley, W. J. (eds.): *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, Menlo Park, CA (1991). (この成書は，*Proc. of 1989 IJCAI Workshop on Knowledge Discovery in Databases*, Detroit, MI (August 1989) を改訂増補して出版されたものである.)
22) Piatetski-Shapiro, G. and Matheus, C. J.: *Knowledge Discovery Workbench: An Exploratory Environment for Discovery in Business Databases*, 文献 19), pp. 11-24 (July 1991).
23) Quinlan, J. R.: *Induction of Decision Trees, Machine Learning*, Vol. 1, No. 1, pp. 81-106 (1986).
24) Schlimmer, J. C.: *Learning Determinations and Checking Databases*, 文献 19), pp. 64-76 (July 1991).
25) Shen, W. M.: *Discovering Regularities from Knowledge Bases*, 文献 19), pp. 95-107 (July 1991).
26) 篠原 歩，宮野 悟：PAC 学習—確率的で近似的に正しい学習，情報処理，Vol. 32, No. 3, pp. 257-263 (May 1991).
27) Silberschatz, A., Stonebraker, M. and Ullman, J.: *Database Systems: Achievements and Opportunities, The "Lagunita" Report of the NSF Invitational Workshop*, TR-90-22, Dept. of Computer Science, Univ. of Texas at Austin (1990).
28) 園生賢一，河野浩之，西尾章治郎，長谷川利治：大規模知識データベースにおけるサンプルから属性指向によって導出されたルールの評価，第5回人工知能学会全国大会論文集，pp. 181-184，東京 (June 1991).
29) 横田一正：演繹オブジェクト指向データベースについて，コンピュータソフトウェア，Vol. 9, No. 4, pp. 285-300 (July 1992).
30) 横田一正，西尾章治郎：演繹・オブジェクト指向データベース，情報処理，Vol. 31, No. 2, pp. 234-243 (Feb. 1991).
31) Zhong, N. and Ohsuga, S.: GLS—*A Methodology for Discovery Knowledge from Databases*, To appear in *Proc. of the 13th International CODATA Conference* entitled "New Data Challenges in Our Information Age", 1992.

(平成4年7月6日受付)



西尾章治郎 (正会員)

1975 年京都大学工学部数理工学科卒業. 1980年京都大学大学院工学研究科 (数理工学専攻) 博士課程修了. 工学博士の学位を取得. 京都大学工学部助手，大阪大学基礎工学部および情報処理教育センター助教授を経て，1992 年より大阪大学工学部情報システム工学科教授 (知識システム工学講座担当) となり，現在に至る. この間，カナダ・ウォータールー大学，ビクトリア大学客員. データベース，知識ベース，分散システムの研究に興味をもっている. 現在，IEEE Transactions on Knowledge and Data Engineering を含む4論文誌の編集委員. ACM, IEEE など6学会各会員.