情 報 学 基 礎 21－1
（1991. 5. 28）

大規模データベースにおける発見 – 研究の動向と展望

西尾 章治郎

大阪大学情報処理教育センター

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”データが充分に解析されないまま，データベースとして格納されている．これら大量データの有効利用を目的として，多様な知識利用に柔軟に応じうる大規模な共用知識ベースを構築することが，今後の重要な研究課題として注目されている．その構築のためには，大規模データベースに内在するルール（知識）を自動的に発見し，その獲得された知識をデータベースに付加して高度な知識ベースを形成する技術の開発が急がれている．本稿では，大規模データベースにおける知識獲得の最近の研究動向と今後の展望に関して述べることにする．

Knowledge Discovery in Very Large Databases – Current Trend and Future Perspective –

Shojiro NISHIO

Education Center for Information Processing, Osaka University
Toyonaka, Osaka 560, Japan

Currently very large amount of “raw” data are stored as a database without their detailed analysis. To utilize these data effectively and respond flexibly to a wide variety of knowledge applications, the development of very large shared knowledge bases is of prime concern. For this purpose, the technique of how to mine “knowledge” from very large databases becomes a very important issue, and by adding such discovered knowledge to the databases we can construct advanced knowledge bases. This paper gives a brief survey on the current trend and future perspective on the research of “knowledge discovery in very large databases”.

1 はじめに

現在，医学，商業・経済学，科学技術，CAD/CAM などさまざまな分野における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られる．なかには，利用者によって与えられる全データの数が数百万から数十億も格納された**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”の根源であり，このような超大量なデータベースを有効利用するためには，データベースから有用な知識（ルール）を迅速かつ的確に検索する必要がある．これらの知識を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（つまり，内包データベース）の装備

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

このような大規模知識ベースの構築を考えるとき，もし，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，データベースが複雑になるにつれ，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，複雑なデータを簡単には処理できなくなる．もちろん，このような人間の能力をそのままコンピュータに移し換え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実現している．たとえば，そのうちの一つに，大豆の病気診断のためのエキスパート・システム PLANT/DS がある．このシステムでは，専門家による診断プロセスの形式化とコンピュータによる診断例からの帰納という二つの方法で診断規則を求めたところ，コンピュータによって帰納的に導出されたルールのほうが専門家によって形式化されたルールよりも成績が良かったという報告がある [7]．また，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の有用な知識を獲得する研究は開始されており（文献 [1, 3] などを参照），実用システムも開発されている（例として，文献 [14] 参照）．さらに，大量データからその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] の 9 章参照）．このような状況下で実用的なルール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充分にあると考えられる．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技術を研究・開発をする絶好の機会を迎えているともいえる．

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発見するために最近提案されたアルゴリズムのうち次の 2 種が顕緒である．一つは，関係データベース (relational database) 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．G. Piatetsky-Shapiro は，組指向のアルゴリズムとして，すべての条件の**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 [9, 10]．もう一方は，オブジェクト指向データベース (object-oriented database) においてモデル化されるような，属性間の階層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，関係データベースの属性に目を向けて，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くようなを方法を提案している [4, 6]．

しかし，全データの数が数百万から数十億にもおよぶような大規模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．このような場合に，データから任意に取り出した**サンプル** (sample) にアルゴリズムを適用してルールを発見することが考えられる．しかし，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．G. Piatetsky-Shapiro は，組指向のアルゴリズムを使っ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統計的な正確さの解析をおこなっている [9, 10]．また，関生らは，属性指向のアルゴリズム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おこなっている [12]．

本稿では，まず2章で，大規模データベースにおける知識獲得において採用されている手法が，従来の人工知能における研究のアプローチとどのような関連をもっているかについて述べる．その後に3章で，属性指向の知識獲得アルゴリズムを紹介し，4章で組指向の知識獲得アルゴリズムについて概略を述べる．5章では，サンプル処理と得られたルールの正確さに関する研究を紹介する．最後に6章では，その他の重要事項および今後の課題について述べる．

2 人工知能のアプローチとして位置づけ

2.1 広く浅い推論

従来の人工知能システムのアプローチを振り返ってみると，次のように大きく二つに分けられる．

1. システムの適用領域を限定し，知識の核に深く迫ろうとするアプローチ
2. システムの適用領域を広くし，大量の知識を広く浅く推論するアプローチ

これまで，人工知能システムとしての成功を納めてきた医療などにおけるエキスパートシステムの開発は，第一のアプローチによるものである．しかし，対象となるシステムが大規模になってきた場合には，そのアプローチのスケールアップが非常に困難となる．そこで，大規模データベースからの知識獲得には通常スケールアップの可能性をもつ第二のアプローチが採用される．しかし，このアプローチに基づく知識獲得技術は，第一のアプローチと比べてあまり進んでおらず，今後の研究・開発が非常に重要視されている [13]．大規模データベースからの知識獲得に対して期待されることは，対象領域が狭くて深い知識より，浅くても良いから大域的に成立するルールであり，そのルールをデータベースに付加することにより，少しでもデータベースを知識化ができ，探索の効率化あるいは再帰的な問合せの評価 [8] などが可能になることである．

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に関して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 [7]．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事実の集合であり，反例が存在しない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なうために，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)」を行なわないような制約を設ける必要がある．

3 属性指向の知識獲得アルゴリズム

J. Han らの提案した属性指向のアルゴリズム [4, 6] では，関係データベースのすべての属性に対して，学習作業に関する知識を概念木として与え，属性ごとにその概念木をのぼることによって一般化をおこない，組の数を減らしていくことによって，ルールが導出される．本章では，例を用いてこのアルゴリズムの簡単な紹介をすることにする．

まず，このアルゴリズムを実行するうえで，データベースの設計者は，属性間の階層構造に関する知識をもちながらデータベースのスキーマの設計をしたという前提がある．つまり，情報工学科，化学工学科，機械工学科などが工学部に属すること，工学部，理学部が理科系に属するというような階層構造に関する知識があるという仮定のもとでアルゴリズムが実行される．

以下では，大学に属する人に関して，属性の組（名前，性別，年齢，住所，所属，身分）からなる関係データとそのデータに関する概念木が与えられているという仮定のもとで，学生に関して成立するルールを導出することを考える．特に，与えられたデータの一部分である（名前，性別，年齢，住所，所属）を学生の属性して考える．

〔一般化の手順〕 まず，**最終段階の一般化関係** (final generalized relation) に残す組の数を，利用者や専門家によって前もって特定されたしきい値 (threshold) とする．ここで示す例では，しきい値を３とする．

手順１（学習作業に関連したデータの選択 (前処理)）

選択，射影，結合などにより，全データから学習作業に関連するデータだけを取り出す．例では，学生についての関係が取り出され表１のように与えられる．

手順２（属性の除去）

もし，属性にしきい値以上の異なる値をもつ大きな属性値の集合があるにもかかわらず，その属性により高度なレベルの概念が与えられていないなら，その属性は一般化の過程において除去する．これは，教示学習の一般化規則の一つである**条件削除規則** (dropping condition)[7] に対応する．

一般化は，全データの属性ごとに順番に行なっていく．そうすると，作業に関連しない属性が存在することがあり得る．例では，最初の属性である名前には，より高度なレベルの概念は特定されていない．一般化において，このような属性は除去すべきである．なぜなら，学生の一般の特性は，属性・名前によって特徴づけられないことを意味するからである．ただし，属性・性別のように，異なる属性値の数がしきい値よりも小さいものは，除去せずに残しておく．したがって，表１から属性・名前が除去された表２が新たに得られる．

名前	性別	年齢	住所	所属
尾崎	男	１８	京都市	物理学科
鬼村	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	法学科
渡部	男	２２	宝塚市	応用化学科

表１：大学に属する学生のデータ

性別	年齢	住所	所属
男	１８	京都市	物理学科
女	２２	相生市	国文学科
男	２３	城陽市	数理工学科
女	２２	西宮市	国文学科
男	２３	西宮市	法学科
男	２２	宝塚市	応用化学科

表２：条件削除規則によって得られるデータ

手順３（概念木の上昇と冗長な組の除去）

もし，ある組の属性値の概念に，より高度なレベルの概念が存在するなら，その値をより高度なレベルの概念で置き換え，冗長となる組を除去することにより一般化できる．一度に１レベルずつ概念木をのぼることにより，一般化しすぎてしまうことを防ぐ．しかも，ある概念と共に，より一般化された概念が共存することを防ぐために，すべての属性値が概念木を同時にのぼる．これは，教示学習の一般化規則の一つである**一般化木上昇規則** (climbing generalization tree)[7] に対応する．

表2の四つの属性（性別，年齢，住所，所属）のそれぞれの値を，より高度なレベルの概念で置き換えることにより一般化をおこなうことができる．たとえば，京都市は京都府にそして関西に … というように一般化できる（もちろん，このような一般化は前提である概念木に基づいている）．そのような置き換えは，属性ごとに順番におこなう．その結果，表3が得られる．

手順4（各属性に対するしきい値の操作）

しきい値よりも作業に関連した属性の異なる値の数が大きければ，この属性に対してさらに一般化をおこなう．例では，所属に関する一般化が行なわれ，その結果の関係が表4で与えられる．

性別	年齢	住所	所属
男	若年	京都府	物理学科
女	若年	兵庫県	国文学科
男	若年	京都府	数理工学科
男	若年	兵庫県	法学科
男	若年	兵庫県	応用化学科

表3：一般化木上昇規則によって得られるデータ

性別	年齢	住所	所属
男	若年	京都府	理学部
女	若年	兵庫県	文学部
男	若年	京都府	工学部
男	若年	兵庫県	法学部
男	若年	兵庫県	工学部

表4：しきい値を操作して得られるデータ

属性ごとの概念木の上昇と冗長な組の除去が終わっても，一般化された関係における組の総数は，しきい値よりも大きいかもしれない．このような場合，さらに一般化をおこなわなければならない．このために，手順5を考える．

手順5（一般化された関係に対するしきい値の操作）

しきい値よりも一般化された関係における組の数が大きければ，この関係に対してさらに一般化をおこなう．

この段階では，さらに一般化をおこなうために候補とする属性の選択がいく通りか考えられる．一般には，組の数や異なる属性値の数の縮小率が大きいものを優先するなどの手段が使われる．ところが，候補とする属性の選択の仕方を変えていろいろなルールを作ると，関心のあるルールが異なる**一般化経路** (generalization path) をたどって発見されることもあり得る．これは，同じデータの集合からでも，人によって学習する内容が違うことがあるということに相当する．したがって，一般化された関係が導出できれば，関心のあるルールだけを残すように，ユーザや専門家によって平凡なものは取り除かれるべきである．例に関しての結果を表5に示す．ただし，ANY（属性）は（）内の属性の最も一般化された概念であることを表わす．

性別	年齢	住所	所属
男	若年	京都府	理系
ANY(性別)	若年	兵庫県	文系
男	若年	兵庫県	理系

表5：最終段階に残ったデータ

最終段階の一般化関係は少数の組だけから成る．従って，簡単な論理式に変型できることから，次の手順6を考える．

手順6（ルールの変型）

最終段階の一般化関係の中の一組は，連言標準形に変型できる．組どうしは，選言標準形に変型できる．

表5の最終段階の一般化関係は，次のように論理式に変型できる．ここでは，一組目だけを明示してある．残りのものについても同様なので，省略する．

$\forall x$ [(身分 (x) $\in$ 学生) $\rightarrow$ ((性別 (x)= 男) $\wedge$ (年齢 (x) $\in$ 若年) $\wedge$ (住所 (x) $\in$ 京都府)$\wedge$ (所属 (x) $\in$ 理系)) $\vee$ $\cdots$]

4 組指向のアルゴリズム

本章では，組指向のアルゴリズムに関して，その考え方のみについて簡単な説明をする．詳細は，文献[9, 10]を参照されたい．

ここで，データベースのデータが，前章で例に用いた表2のように与えられていたとしよう．まず，性別に注目し，「男」と「女」によってすべての組を二つに分類する．「女」の組に注目して，次に年齢の属性を調べるとすべて「22歳」であることが分かる．さらに，所属に注目するとすべて「国文学科」であることが分かる．しかし，住所に関しては異なった値をとっている．以上の観察から，与えらたデータに関して，「女学生であれば，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することが分かる．

G. Piatetsky-Shapiroは，以上のような知識獲得の手順を一般化し，さらに，計算効率を考慮してアルゴリズムを記述している．特に，データベースのある属性について，各組へのアクセスを1回しか要求しないように，属性値に応じてすべての組をセル(cell)にハッシュ(hash)する手法を採用している．このようにして，ある属性について，同じ属性値をもつ組だけをデータベースから選び出して，一つのセルとして分類し，そのセルの中で，他の属性について，すべての属性値が同じであるような属性が存在すれば，ルールが導出できる．これは，属性値が同じになるような組をデータベースから選び出すので，組指向のアルゴリズムといえる．

この手法の利点として，ルールの導出に関しては各セルは独立にアルゴリズムを実行することができ，並列処理が実現し易いことである．この性質は，対象となるデータベースの規模が大きくなればなる程，非常に有効なものとなる．

5 サンプル処理と評価

大規模データベースにおける全データは数百万から数十億あるため，完全なルールを導出するには非常に効率の良いアルゴリズムを用いても困難であろう．このような場合に全データから任意にサンプル抽出を行ないルールを導出することが有効と考えられる．しかしながら，サンプルから導出されたルールがデータベースの全データにおいて妥当であるとは限らない．そこで，大規模データベースの全データから導出されたルールとサンプルから導出されたルールを比較してその正確さを評価しておく必要が生じる．これまでに属性指向のアルゴリズムを用いたサンプルからのルールの導出とその正確さの評価はG. Piatetsky-Shapiro[9, 10]によって行なわれており，属性指向のアルゴリズムの場合のサンプル処理とルールの正確さの評価は園生ら[12]によって行なわれている．

本章では，サンプル抽出を行なう属性指向のアルゴリズムに関して得られた結果を簡単に紹介することにする(サンプル抽出を行なう属性指向のアルゴリズムと評価式の導出過程は，文献[12]に詳しい)．

まず，条件 $A = a$ が与えられたとき，大規模データベースの全データから任意に抽出されたサンプルから，

$$(A = a) \rightarrow (B = b_1 \vee b_2 \vee \cdots \vee b_{t_s}) \tag{1}$$

というルールが導出されたとする．ここで，A, Bは属性で，$a, b_i (1 \leq i \leq t_s)$ は属性値である．

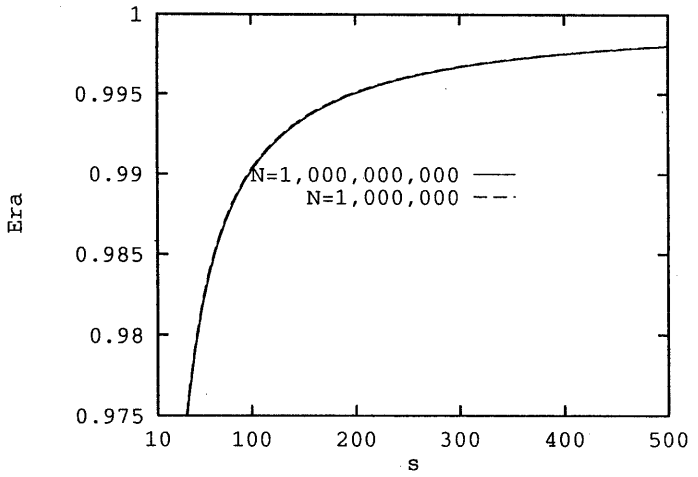

図 1: 条件 $A = a$ を満たす組の数 s に対するルールの正確さ E_{ra}

この時，全データから導出されたルールに対するサンプルから導出されたルールの正確さ E_{ra}の上限・下限が次のように与えられる．

$$E_{ra}(n) \approx 1 - \frac{1-s/n}{s+2} + \frac{(s+1)\frac{(1-\frac{t_s}{2})\frac{\lambda^2}{s}+\sum_{i=1}^{t_s}\sqrt{\frac{\lambda^2}{s}\hat{p}_i(1-\hat{p}_i)+\frac{(\frac{\lambda^2}{s})^2}{4}}}{1+\frac{\lambda^2}{s}}e^{-Y}}{(s+2)(1-e^{-Y})}$$

ここで，

$$Y = \frac{(s+1)n\frac{(1-\frac{t_s}{2})\frac{\lambda^2}{s}+\sum_{i=1}^{t_s}\sqrt{\frac{\lambda^2}{s}\hat{p}_i(1-\hat{p}_i)+\frac{(\frac{\lambda^2}{s})^2}{4}}}{1+\frac{\lambda^2}{s}}}{[n\{2-\frac{(1-\frac{t_s}{2})\frac{\lambda^2}{s}+\sum_{i=1}^{t_s}\sqrt{\frac{\lambda^2}{s}\hat{p}_i(1-\hat{p}_i)+\frac{(\frac{\lambda^2}{s})^2}{4}}}{1+\frac{\lambda^2}{s}}\}+3]/2}$$

$$N\frac{\frac{s}{S}+\frac{\lambda^2}{2S}-\sqrt{\frac{\lambda^2}{S}\cdot\frac{s}{S}(1-\frac{s}{S})+\frac{(\frac{\lambda^2}{S})^2}{4}}}{1+\frac{\lambda^2}{S}} < n < N\frac{\frac{s}{S}+\frac{\lambda^2}{2S}+\sqrt{\frac{\lambda^2}{S}\cdot\frac{s}{S}(1-\frac{s}{S})+\frac{(\frac{\lambda^2}{S})^2}{4}}}{1+\frac{\lambda^2}{S}}$$

である．

図 1 には，サンプル $\mathbf{S}$ の中で条件 $A = a$ を満たす組の数 s を 10 から 500 まで変化させた場合のルールの正確さ E_{ra}の下限値を示した．この時，全ファイル $\mathbf{F}$ 中の組の総数 $N = 10^9, 10^6$， サンプル $\mathbf{S}$ 中の組の総数 $S = 10^4$， 最終段階で一般化関係に残されたしきい値に相当する組の数 $t_s = 10$，信頼区間 95% ($\lambda = 2$)，データの分布は最良の場合を仮定して標本比率 $\hat{p}_i = 0.1(1 \le i \le 10)$ とした．また，他の変数を固定したまま $\hat{p}_i = i/55(1 \le i \le 10)$ と変化させた場合，更に $t_s = 3, \hat{p}_i = 1/3(1 \le i \le 3)$ と変化させた場合も上と同様の結果を得た．

以上のような結果により，サンプル抽出を行なう属性指向のアルゴリズムによって導出されるルール (1) の正確さはデータの分布$\hat{p}_i$やしきい値 t_sや全ファイル中の組の総数 Nにはほとんど影響せず，条件 $A = a$ を満たす組が一定以上の値になれば，比較的少ないサンプルでルールの正確さが 1 に急速に収束すること

が明らかにされた．また，ルールの正確さは条件 $A = a$ を満たす組の数 s が同じならば，全ファイル $\mathbf{F}$ 中の組の総数 N が大きい方がルールの正確さがわずかに低くなることも示された．したがって，大規模データベース全体に含まれる組の数に比べて非常に少ないサンプルによって十分正確なルールが導出できるため，全データから任意にサンプル抽出を行なう属性指向のアルゴリズムは，大規模データベースにおける知識獲得において有効であることが示された．

6 今後の課題と展望

本稿では，「大規模データベースにおける発見」と題して，データベースにおける知識獲得に関する話題を中心として述べた．本稿のまとめとして，この分野で今後重要になると思われる研究テーマのうち三つについて述べることにしよう．

より高性能なアルゴリズムの要求

大規模データベースからの知識獲得における最重要な要件は，計算時間の爆発を防ぐためのアルゴリズムの効率的な実行である．その解決策として，4 章でも述べたようなアルゴリズムの並列実行と 5 章での論じたサンプルからの知識獲得は有効な方法であり，今後さらにいろいろな側面からの研究が待たれる．

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人々の間では，「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規則性に従わないたったの 1 割のデータ（ノイズ）のために，全体の管理に要する時間の 9 割を要しているいうことである [5]．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ルールがデータの更新後満たされなく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象する技術としては，次に列挙する方法が考えられる [5]．

- 例外処理 (exception handlers)
- ルールの動的改訂
- 再定義 (overriding)
- 不履行 (defaults) 論理
- 異常宣言述語 (abnormality predicates)
- 確率の導入．計算の複雑度と関連した PAC(probably approximately correctly learnable) 学習 [11]．
- ファジィー論理

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．

マルチメディア情報の知識ベース化

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．

謝辞 本稿をまとめるにあたって貴重なコメントを頂いた京都大学工学部数理工学教室河野浩之氏に衷心より感謝の意を表わす．なお，著者の本稿の分野における研究の一部は，文部省科学研究費によっている．

参考文献

[1] C.S. Ai, P. Blower, and R. Ledwith, "Discovering Reaction Information from Chemical Databases", *Proc. IJCAI-89 Workshop on Knowledge Discovery in Databases*, Detroit, MI, August 1989.

[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, 岩波書店, 1989.

[3] R. Blum, "Discovery and Representation of Causal Relationships from a Large Time-Oriented Clinical Database: the RX Project", *Lecture Notes in Medical Informatics*, No.19, Springer-Verlag, 1982.

[4] Y. Cai, N. Cercone, and J. Han, "An Attribute-Oriented Induction in Relational Databases", *Proc. IJCAI-89 Workshop on Knowledge Discovery in Databases*, Detroit, MI, August 1989.

[5] C. Esculier, "Non-Monotonic Knowledge Evolution in VLKDBs", *Proc. of 16th Very Large Data Bases*, Brisbane, Australia, pp.638-649, August 1990.

[6] J. Han, Y. Cai, and N. Cercone, "Data-Driven Discovery of Quantitative Rules in Relational Databases", To appear in *IEEE Transactions on Knowledge and Data Engineering*, 1991.

[7] R. Michalski, "A Theory and Methodology of Inductive Learning", in R. Michalski, et al. (eds.), *Machine Learning: An Artificial Intelligence Approach, Vol.1*, Morgan Kaufmann, 1983.

[8] 西尾, 楠見, "演繹データベースにおける再帰的な問い合わせの評価法", 情報処理, Vol.29, No.3, pp.240-255, 1988.

[9] G. Piatetsky-Shapiro, "Discovery and Analysis of Strong Rules in Databases", *Proc. IJCAI-89 Workshop on Knowledge Discovery in Databases*, Detroit, MI, August 1989.

[10] G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules", To appear in P. Piatetsky-Shapiro and W. Frawley (eds.), *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991.

[11] 篠原, 宮野, "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, 1991.

[12] 園生, 河野, 西尾, 長谷川, "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, 掲載予定, 1991.

[13] 横井俊夫, "知識の知識ベース", 大規模知識ベースに関するシンポシウム予稿集, (財) 京都高度技術研究所, pp.3-7, 1990.

[14] W. Ziarko, "Data Analysis and Case-Based Expert System Development Tool ROUGH", Proc. of Workshop on Case-Based Reasoning, Pensacola Beach, FL, Morgan Kaufmann, 1989.