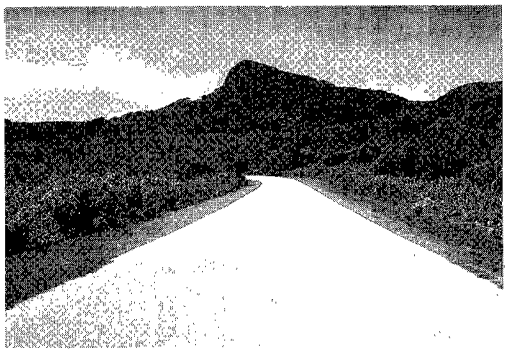
道しるべ：データ活用の礎 —データウェアハウスとデータマイニング—



河野 浩之

京都大学大学院 情報学研究科
システム科学専攻

カセットテープレコーダーなどからコンピュータプログラムやデータを読み出した経験がある者には，今は夢のような時代になった．なんといっても，計算機性能の著しい性能向上がもたらした世界の広がりは大きな刺激を与えてくれる．もっとも，紙テープやパンチカード以前の時代を経験された方からは，まったく別の意見が出そうな気もする．「確かに昔は豊かではなかったが，プログラムとしてもデータとしても利用できるようにメモリをやりくりし，1ビットすら無駄にせず大切に利用したものだ．それにひきかえ，今の時代の若者は…」などと続くかもしれない．

しかし，あり余るほどの広さを手中にすれば，極限まで切り詰めたスタイルをとらない人が多くなるのは世の常だ．そこで，このデータの溢れる時代ならではの，データ活用を彩る研究を眺めていくことにしよう．

データベースの歩み

まず，コンピュータを取り巻く資源そのものに余裕がなかった時代からデータベース技術がどのように変化してきているのかを振り返っておこう[8), 11)]．

データベースは，データを正確に記録し，実際に役立つ処理を行うことが求められる．だからこそ，実体（entity）と関連（relation）の在り方を考えながら慎重に世界を削り出し，必要不可欠な属性（attribute）からなるスキーマを設計し，無駄のないデータの記録を行う．そして，インデックス技術とデータ構造を駆使した高速なデータ検索技術や，複数ユーザによるデータ書込／読出／更新／削除を誤りなく実行するトランザクション処理技術の開発が，データベースシステムの基礎となっている．その結果，金融トランザクション処理，ビジネスデータベース，エンジニアリングデータベースなど数多くのデータベース応用システムとして社会へと受け入れられた．また，その成功が，より高速なトランザクション処理と大量データ検索を実現する現在の大規模データベース構築へとつながった．

ところで，現在よく利用されている関係データベースを例にとると，属性間の関係を示す関数従属性を利用しながらテーブル作成を行う正規形の考え方がスキーマ設計に必要である．もちろん，このような正規形を最大限に利用することで，データの冗長性を減らし，データの一貫性を保証するという大きなメリットを得られるからこそ，厳しい制約が当然のものとして受け入れられてきた．

ただし，ここで注意しておかねばならないのは，データベースが蓄積可能なデータ量は増大したが，記録するデータに対する質的な制約は依然として厳しいものを要求し続けていることだ．つまり，不完全でノイズがあるような大量の生データを，現在利用されているデータベースに蓄積しても，思うような処理ができないのは当然のことである．

しかし，このような制約が逆に束縛となり，本当に必要なデータ蓄積や処理や検索ができなくなってしまうことは，データを活用するうえで最も避けなければならない．また，新しいタイプのデータ活用を行いたいという要求は強く，たとえば，意思決定支援に必要な実時間データ解析，ネットワークで接続された複数のデータベースに対する連携的な検索，さらに，曖昧なデータに対する探索的処理や特徴発見など多数ある．

そこで，大量の生データを手にし，新たな処理方法が求められている現在，データ活用の礎を築くために，現実のデータに近い形で種々の問合せ処理を実現する研究が行われている．以下，いくつかのチェックポイントを通過しながら進むことにする．

企業データ活用の鉄則

データベースが，企業内の業務処理で大きな市場を獲得してきたという点に重きを置くなら，これまで築き上げられた大量のデータベース資産を，どのように活用すべきかということは，やはり市場の視点から考えておかねばならない．

したがって，まずデータ活用技術の第1のチェックポイントとして取り上げるべき項目は，エンドユーザ向けの意思決定支援システム（DSS; Decision Support System）として位置付けられるデータウェアハウス（data warehouse）[6]である．

データウェアハウスの概念を提唱したInmon氏は，「サブジェクト指向」「統合化」「時系列」「恒常性」を持つ情報システムとしてデータウェアハウスを位置付けている．もっとも，これらを満たすシステム構成技術は互いに強く関係しているので，ここでは，次のような角度から整理しておく．

まず，業務処理の終了したデータを廃棄したり，ただ，テープ媒体に眠らせておくのでは，データ資産の活用などできない．そこで，業務上利用したデータを削除することなく蓄え，1週間や1月といった時間間隔で要約し格納する．もちろん，支店や支部といった組織に合わせた要約の格納も必要である．なお，このような処理を自然に実現するには，正規形に拘束された関係データベースのスキーマ設計に拘ることなく，スタースキーマ（star schema）やスノーフレークスキーマ（snowflake schema）と呼ばれる柔軟なスキーマ設計が役立つ．

さらに，意思決定を行う場面に応じて，異なる時間間隔でのデータ解析，異なる組織の切口でのデータ要約を瞬時に行わねばならないこともある．そこで，多量のデータを実時間処理するために，関係データベースを操作するのではなく，多次元データベース（multidimensional database）へとデータを移動して処理速度を向上させる設計がとられる．なお，このような解析処理を実時間で行うOLAP（On-Line Analytical Processing）技術を，第2のチェックポイントとしておく[5]．

また，データウェアハウスでは，特定のテーブルに対する接合（join）演算を頻繁に必要とする問合せも多い．ここでは，問合せの前処理結果をデータベースに格納しておくことが，最も自然なシステム設計となるだろう．そして，このような前処理を効率よく行う研究は，データウェアハウスを理論付けるものとされ，たとえば，データキューブ（data cube）として知られている．なお，この種の研究を大きく捉えると，データ格納に要する記憶領域を最小にするのではなく，問合せ処理と記憶領域の両面からコスト評価を行うデータベース設計技術であると考えられる．そこで，ビュー実体化（materialized view）技術を第3のチェックポイントとしておく☆1．

ところで，データ活用を進めるうえで，企業内で利用されている数多くのデータベースを連携しながらデータ統合を目指すことも当然必要である．また，企業の状態を総合的に把握できるデータを整備することは，意思決定における最も基本的な姿勢でもある．もっとも，多数の業務データベースにおけるデータ型，値域，属性値の記述方法などを統一することは非常に困難であると予想される．そこで，すべての業務プロセスを効率的に実現できるような設計図をトップダウン的に与えるERP（Enterprise Resource Planning）がある．逆に，複数データベースの連携統合をボトムアップ的に支える技術として，ラッパー（wrapper）やメディエータ（mediator）と呼ばれる技術がある．後者を第4のチェックポイントとしておきたい[2], ☆2.

データから情報を取り出すには

データ活用技術として注目しておかねばならないのが，データベースからの知識発見（KDD; Knowledge Discovery in Databases）やデータマイニング（データ発掘）（data mining）と呼ばれる研究である．

特に，データマイニングでは，OLAPのような解析処理以上に，より高度な問合せ処理を要求している点に注意しておきたい．たとえば，「頻繁に組み合わせて購入される商品は何か？」「一定区域の中で最も若者が密集している場所はどこで，その時間帯は？」「Webページでバナー広告表示するときに，利益率の最も大きくなるような動的リンク生成ができるか？」さらには，「与えられたデータの性質を最もよく表現したルールはどのようなものか？」などといった問合せである．これらの問合せを行う記述を，果たしてSQLのような標準的な問合せ言語を用いて書けるだろうか．そして，もし書けるとしても，実用上の問題は生じないだろうか．

まず，正確な問合せ記述を与えることは面倒ではあるが，本質的な処理技術はすでに確立されていると判断する人もいるかと思う．ところが，現実のデータ処理を行うと，行き詰まることが多い．実際，単調な問合せを何度も繰り返して目的の解を得られなかった検索を行った人も多いだろう．したがって，標準的問合せ言語程度で処理可能な範囲にあるという判断は正しいが，飽き飽きするほど似通った問合せを繰り返す必要がある点で，データ活用を行う道具としては不十分である．

そこで，過去，この種のデータ探索を行う処理ツールが数多く開発されてきた．また，現在，複数の処理機能を統合化しながら知識発見ツール（KDDM Tools; Knowledge Dis-

☆1 http://www-db.stanford.edu/pub/のJeffrey Ullman's papersから“Implementing Data Cubes Efficiently”などにアクセスできる．
☆2 文献9）の第16章“Currnet Issues（pp.571-603）”に幅広い問題が一覧されている．

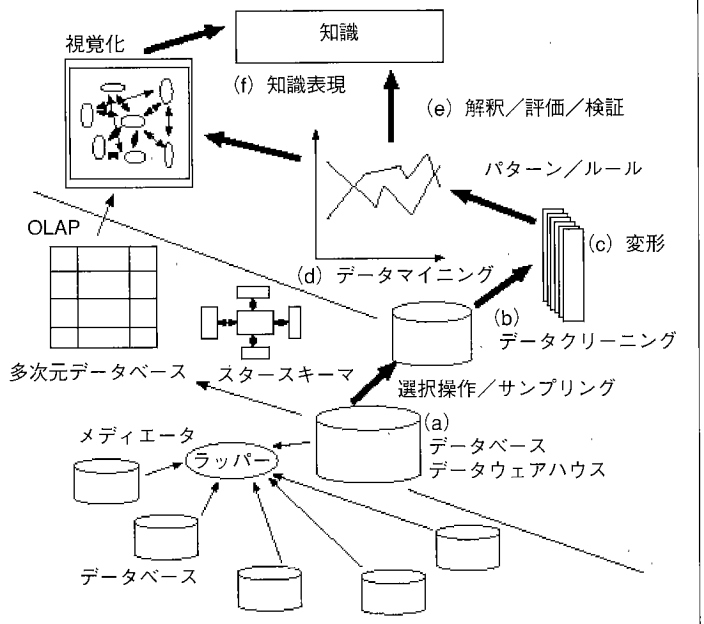


図-1　知識発見プロセス

covery and Data Mining tools）の開発が活発に進められている．そして，データマイニングがツール開発を積極的に推進するからこそ，データ活用の礎を築くうえで重要となり得ていることを考え，これをもって第5のチェックポイントとしておく．

ところで，図-1には，これまで述べたことを踏まえて，知識発見の典型的な過程[4]を簡単に示した．図-1の全体の構成は自然なものであり，以下に述べるステップを見通しよく処理するツールが，データ活用を促進することにも納得できるだろう．

【知識発見プロセス】
(a) データの特性を理解したうえで，データベース／データウェアハウスを構築する．
(b) 選択操作／サンプリング（selection/sampling）によりデータクリーニング（data cleaning）を行う．
(c) 処理可能なデータにするための前処理／データ変換（preprocessing/data transformation）を行う．
(d) データマイニングを行い，ルールを求める．
(e) ルールの解釈／評価（interpretation/evaluation）を行い，データと検証（verification）する．
(f) ルールを知識とする．

なお，ステップ（d）の部分が，「データを情報へと変え，そして，知識を発見する」という知識処理技術の根幹部分であることから，多くの研究者の興味を集めている．たとえば，「確実性の高いパターン」「妥当な法則」「顕著な性質」「有効性の高いパターン」「理解しやすい規則」「興味深い記述」を求めるために，エントロピーや記述長や複雑さなどに注意して，「短くバランスの良い記述」を高速に求める多数のアルゴリズムが研究されている．もちろん，これらの研究が，第6チェックポイントになる．

しかしながら，データマイニング技術を支える理論は，データベース，機械学習（machine learning），統計，情報の視覚化（information visualization）など数多い．さらに，問題へのアプローチがきわめて幅広いため，ここでは文献3），4），1）を示すにとどめておく．

また，改めて述べるまでもないが，たとえ優れたアルゴリズムであっても，構成しやすいハードウェア上で動かなければ利用できない．したがって，現時点で実装するならば，$N \cdot \log N$程度の計算量を持つアルゴリズムを用いるのが妥当な範囲と考えられていることに注意しておく．もちろん，将来，革新的な計算原理が実用化されれば，バラエティに富んだアルゴリズムが利用できる，より豊かな時代の幕が開くに違いない．

知識メディアを目指して

ところで，個人レベルで得られる最大のデータ源の1つは，間違いなくインターネットを通じて得られる大量のWebページだろう．さらに，Webアプリケーションは，情報共有を推進するうえで役立ち，優れた知識メディアの一角に位置付けられていくと考えられる．しかし，現状のWebページは確かにデータではあるが，情報と呼ぶのはかなり辛いといえる．

そこで，データベースやデータウェアハウス，そして，データマイニングの技術が，大量のWebデータ処理に役立つチャンスが生まれる．すでに，多くの要素技術は取り上げてきたので，ここでは組み合わせながら話を進める．

現在，多くのWebサーバは，Webページをhttpで転送するだけであるが，ページ間を一貫して管理するデータベースとしての機能も求められている．つまり，URLをキーとして，データとして格納されたページを転送するのではなく，要素データからページを動的に生成して転送する機能が必要となる．なお，この機能は，データベース技術をWebサーバに連携させながら実現されつつある．

次に，Webページはまったく独立した組織や個人で作成されており，異なるWebサーバ間で情報共有を行うことは難しくなる．そこで，データウェアハウスの項で触れた，メディエータやラッパーの技術を利用することが考えられる．また，この種の統合を実現するうえで，独立にWebページの作成を行いながら，データ構造を緩やかに与えるXML関連技術[12]が役立つ．なお，この種の緩やかなデータ構造は，半構造データ（semi structured data）と呼ばれており，サーベイ論文[10]が参考になる．

また，膨大な数のWebページの効率よい収集，テキストデータやリンクデータの特性の探索，加えて，Webサーバへのアクセス履歴，Webサーチエンジンへの問合せ履

歴，Webアプリケーションの操作履歴など，Webアプリケーションを通じて得られる膨大な対話的データに対して，テキストマイニング（text mining）やWebマイニング（Web mining）と呼ばれるデータマイニング技術が必要になっている．

参考書，国際会議，国内研究会など

今回は，データベースと人工知能の両面から眺めてきた．簡単にいえば，著しい計算機性能の向上がもたらした世界の変化を，データベースのシステム設計技術と人工知能のアルゴリズムによって捉え直し，互いの研究を実社会で受け入れられる技術とするには，どうすればよいかを探っている状況である．

よって，データベース領域から関連研究を見た場合，VLDB, PODS, ICDEなどのデータベース関連の国際会議などで扱われる．もちろん，それ以外にも多数の会議やワークショップが開催されており，それぞれにおいて，データウェアハウス，データマイニング，Webアプリケーションシステムなどが幅広く取り上げられている．また，国内では，情報処理学会データベース研究会（DBS）や，電子情報通信学会データ工学委員会（DE）が，データベースシステムの研究を中心に扱っている．なお，文部省特定領域研究の高度データベース[7]のページから関連研究を探すことができる．

他方，人工知能領域から見るならば，KDD，PKDD，PAKDDなどのデータマイニング関連の国際会議を挙げることができる．北米を中心としたデータマイニングの研究が最も活発であり，KDDはACM SIGKDD（Special Interest Group on Knowledge Discovery & Data Mining）の国際会議となっている．ただし，国際会議となるまで，AAAIの併設ワークショップとして開催されていた．また，国内では，ソフトウェア科学会データマイニング研究会（DM）や人工知能学会知識ベース研究会（KBS）を中心に活発な研究が行われている．さらに，各種学会のチュートリアルテーマとしてデータマイニングが取り上げられたり，各種知識発見ツールのコンテストなども行われている．なお，文部省特定領域研究の発見科学（discovery science）のページから関連研究を見つけることができる．

ところで，本稿では，今回取り上げたテーマに対して初心者が興味を持てる道標を示すため，手軽に入手できる文献を中心に示した．もっとも，ここに至るまでに述べたいくつかのチェックポイントは，現在活発に研究されているテーマであり興味深い文献も多い．そこで，stanford.edu, kdnuggets.com, db.cs.sfu.ca, w3c.orgなどで提供されるWebページを頼りに先へと進むことを期待して，関連するURLを表-1に示しておく．

項目	URL
データウェアハウス[2]	http://www-db.stanford.edu/warehousing/
高度データベース[7]	http://banjo.kuis.kyoto-u.ac.jp/juten/
KDD関連[4]	http://www.kdnuggets.com/
KDDMツール関係[5]	http://db.cs.sfu.ca/DBMiner/survey/table.html
XML-QL[12]	http://www.w3.org/TR/NOTE-xml-ql/
XQL[12]	http://metalab.unc.edu/xql/

表-1　データ活用技術関連のURL

新しいデータ処理への道程

最後になったが，データベース開発者たちは，システム性能を劣化させる可能性を持つ技術の実装を非常に慎重に避けてきた．そのため，新たなデータ活用技術の原型となるアイデアは過去の研究の中にも見つかるのだが，実を結ぶまでには至っていない．

しかし，現在，新たなタイプのデータ活用技術を備えたデータベースの形がようやく整いつつある．もっとも，さまざまな技術をバランスよく鍛え上げるまでには，まだ時間を要すると思われる．だが，今後も，多様化するメディアによりデータ量は単調に増大する一方である．時代にマッチしたデータ活用法をしっかりと考えておくことは，さまざまな情報システムにおいて効果的なデータベース利用を行ううえで必ず役立つに違いない．

参考文献

1) Berthold, M. and Hand, D.J. (ed.): Intelligent Data Analysis, An Introduction, Springer (1999).
2) IBM Corporation: The Garlic Project (1996). http://www.almaden.ibm.com/cs/garlic/homepage.html
3) Cabena, P. et al.: Discovering Data Mining, Prentice Hall (1998). キャベナ他 (著), 日本IBM, 河村, 福田 (監訳): データマイニング活用ガイド, トッパン (1999).
4) Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press (1996).
5) Han, J., Lakshmanan, L.V.S. and Ng, R.T.: Constraint-Based, Multidimensional Data Mining,COMPUTER (special issues on Data Mining), Vol.32, No.8, pp.46-50 (1999).
6) Inmon, W.H.: Building the Data Warehouse, John Wiley & Sons (1996). インモン: データウェアハウス構築編, オーム社 (1997).
7) 上林弥彦 他: データベース研究21世紀への提言 (文部省特定領域研究「高度データベース」とデータベース研究の今後), Computer Today, No.89, pp.33-37 (1999). (No.89より7回にわたる連載).
8) Kim, W. (ed.): Modern Database Systems, ACM Press, New York (1995).
9) Özsu, M.T. and Valduriez, P.: Principles of Distributed Database Systems, Second Edition, Prentice Hall (1999).
10) 田島敬史: 半構造データのためのデータモデルと操作言語, 情報処理学会論文誌データベース, Vol.40, No.SIG 3 (TOD 1), pp.152-170 (1999).
11) Ullman, J.D.: Principles of Database Systems, Computer Science Press, Inc. (1982).ウルマン (著), 國井, 大保 (共訳): データベース・システムの原理, 日本コンピュータ協会 (1985).
12) W3C Technical Reports and Publications. http://www.w3.org/TR/

(平成11年11月2日受付)